Damian Borth

Visual Learning of Socio-Video Semantics

Dissertation

genehmigt vom Fachbereich Informatik der Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

Dekan:
Prof. Dr. Klaus Schneider, Universität Kaiserslautern

Berichterstatter:
Prof. Dr. Thomas Breuel, Universität Kaiserslautern
Prof. Dr. Andreas Dengel, Universität Kaiserslautern

Vorsitzender der Promotionskommission:
Prof. Dr. Hans Hagen, Universität Kaiserslautern

Datum der Aussprache:
11. Juli 2014

D 386

Für meinen Großvater Paul Rzany

Acknowledgements

I would like to express my gratitude to all those who have helped and supported me during the completion of this thesis. First, I would like to thank Prof. Thomas Breuel for his trust and guidance throughout the time of my dissertation research. I also want to thank Prof. Andreas Dengel, who provided comments, suggestions, and supported my research beyond the scope of this dissertation. I am also grateful to Prof. Shih-Fu Chang for the opportunity to visit his lab and to collaborate on visual sentiment analysis, and to Prof. Klaus Madlener for his support during my early years in academia. In particular, I would like to thank Adrian Ulges, without his mentorship and constant support this dissertation would not have been possible.

Next, I would like to thank all my colleagues at MADM and IUPR for all their fruitful discussions, pointers to related work, and answers to technical questions. This includes in particular colleagues and students who either contributed with their work to this thesis or reviewed it: Jörn Hees, Tim Althoff, Adrian Ulges, Christian Schulze, Christian Kofler, Armin Stahl, Joost van Beusekom, Markus Koch, Alexander Arimond, Marco Schreyer, and Jane Bensch.

An dieser Stelle möchte ich mich auch bei all denen bedanken, welche dafür gesorgt haben, dass ich das Wesentliche nicht aus den Augen verloren habe: Myriam für ihre Unterstützung und Gedult, meine Brüder Arthur und Waldemar für die Erinnerung, dass es ein Leben ausserhalb der Forschung gibt, und meine Eltern für ihre Hilfe und ihr Vertrauen während dieser Zeit.

Abstract

Today's ubiquity of visual content as driven by the availability of broadband Internet, low-priced storage, and the omnipresence of camera equipped mobile devices conveys much of our thinking and feeling as individuals and as a society. As a result the growth of video repositories is increasing at enourmous rates with content now being embedded and shared through social media. To make use of this new form of social multimedia, *concept detection*, the automatic mapping of semantic concepts and video content has to be extended such that concept vocabularies are synchronized with current real-world events, systems can perform scalable concept learning with thousands of concepts, and high-level information such as sentiment can be extracted from visual content. To catch up with these demands the following three contributions are made in this thesis: *(i)* concept detection is linked to trending topics, *(ii)* visual learning from web videos is presented including the proper treatment of tags as concept labels, and *(iii)* the extension of concept detection with adjective noun pairs for sentiment analysis is proposed.

In order for concept detection to satisfy users' current information needs, the notion of fixed concept vocabularies has to be reconsidered. This thesis presents a novel concept learning approach built upon dynamic vocabularies, which are automatically augmented with *trending topics* mined from social media. Once discovered, trending topics are evaluated by forecasting their future progression to predict high impact topics, which are then either mapped to an available static concept vocabulary or trained as individual concept detectors on demand. It is demonstrated in experiments on YouTube video clips that by a visual learning of trending topics, improvements of over 100% in concept detection accuracy can be achieved over static vocabularies (n=78,000).

To remove manual efforts related to training data retrieval from YouTube and noise caused by tags being coarse, subjective and context-depedent, this thesis suggests an automatic *concept-to-query mapping* for the retrieval of relevant training video material, and *active relevance filtering* to generate reliable annotations from web video tags. Here, the relevance of web tags is modeled as a latent variable, which is combined with an active learning label refinement. In experiments on YouTube, active relevance filtering is found to outperform both automatic filtering and active learning approaches, leading to a reduction of required label inspections by 75% as compared to an expert annotated training dataset (n=100,000).

Finally, it is demonstrated, that concept detection can serve as a key component to infer the sentiment reflected in visual content. To extend concept detection for sentiment analysis, *adjective noun pairs* (ANP) as novel entities for concept learning are proposed in this thesis. First a large-scale visual sentiment ontology consisting of 3,000 ANPs is automatically constructed by mining the web. From this ontology a mid-level representation of visual content – SentiBank – is trained to encode the visual presence of 1,200 ANPs. This novel approach of visual learning is validated in three independent experiments on sentiment prediction (n=2,000), emotion detection (n=807) and pornographic filtering (n=40,000). SentiBank is shown to outperform known low-level feature representations (sentiment prediction, pornography detection) or perform comparable to state-of-the art methods (emotion detection).

Altogether, these contributions extend state-of-the-art concept detection approaches such that concept learning can be done autonomously from web videos on a large-scale, and can cope with novel semantic structures such as trending topics or adjective noun pairs, adding a new dimension to the understanding of video content.

Contents

List of Figures

List of Tables

This work was supported by the PhD Program of Computer Science at the University of Kaiserslautern.

Chapter 1

Introduction

Currently, traditional media is experiencing a major shift towards *social media*. In the same way interaction in social media is to an increasing degree enriched with images and videos. This combination gives rise to a new type of content being coined as *social multimedia*. Examples illustrating this development are the Arab Spring in the Middle East in 2012, where public protests were organized, communicated, and propagated by means of social media, or the Boston Marathon Bombings on April 15^{th}2013, where the majority of media coverage – including official sources such as the police and the FBI – was distributed by social platforms like Twitter and YouTube instead of through traditional media. One triggering key element of this trend is the upload and distribution of images and videos over the Internet minutes after such incidents happen. This is possible because of the availability of broadband Internet, low-priced storage, and the omnipresence of camera equipped mobile devices allowing people to record, publish, share, and consume digital images and videos without effort. This ubiquity of visual content conveys much about our thinking and feeling as it reflects our personal life and ourselves as a society.

As the world is turning towards visual communication [SW09], the sizes of image and video databases are growing with enormous rates. Nowadays, users are generating large amounts of video material and are publishing it online via video platforms like "YouTube", "Vimeo", or "Dailymotion" [1]. YouTube, as the most prominent provider in this area, stores about 100 hours of new video content every minute on its database and delivers over six billion hours of videos to its users every month [YOU13]. Additionally, live-streaming services like "Ustream" or "Justin.tv" [2] form another quickly growing area for digital video broadcasting. Besides web video sharing platforms and live streaming services a third form of digital video is moving towards the Internet: streamed video on demand. Platforms such as "Netflix" or "Amazon Instant Video" [3] are starting to conquer this market and they are followed by traditional TV broadcasters like News Corp. and Time Warner. In contrast to these commercial consumer efforts digital video streams are recorded and stored in other contexts including efforts to digitalize a nation's broadcast archives [HSdRS12], to preserve cultural heritage [PKA+07], to maintain public safety by surveillance camera monitoring [BBC06], and to augment reality or record our surroundings[4].

Summarizing, there is not only an immense amount of digital video already stored digitally, but

[1] www.youtube.com, www.vimeo.com, www.dailymotion.com

[2] www.ustream.com, www.justin.tv

[3] www.netflix.com, www.amazon.com/Instant-Video/

[4] www.google.com/mobile/, www.google.com/glass/

also a rapidly growing trend to make even larger quantities of video content available online. This is particularly reflected in a recent report by CISCO [Inc12], according to which global internet video traffic will account for 55% percent of all consumer Internet traffic (excluding P2P traffic) in 2016.

1.1 Video Retrieval

Unfortunately this content is of no use if not made accessible by system providing means to search in it. According to Jain and Hampapur [JH94], the purpose of digital video is to entertain (e.g. TV shows, music videos), to inform (e.g. news broadcasts, documentaries), to communicate (e.g. video conferencing), and to analyze (medicine, surveillance). Considering each of these areas, different types of retrieval mechanisms are required. In the literature three major groups of access mechanisms to visual databases can be found: "query-by-sketch" [SC97, CDBP99, Ege97] or "query-by-example" [FSN+95, NBE+93, BSUB08, dRSW08], where a sketch or an image is given as an example and similar images are returned from the database, and "query-by-text" (also referred to as "textual-search") [NBS+02], where the user formulates a textual query and the retrieval system returns images or videos stored in the database that are associated with the given keywords. While the first two query approaches might favor a browsing kind of exploration of the database, the latter one is considered to be a more natural querying mechanism to the user [YH07] and is therefore used almost exclusively in the context of search engines (it is also the preferred query mechanism for video platforms like YouTube). This thesis will focus on "textual-search" driven video retrieval.

However, to employ "textual-search" an index containing the mapping between keywords and videos must be built. This requires the labeling of the audio-visual video stream by keywords describing its content. According to Snoek and Worring [SW09] two types of semantic labeling approaches exist to build such an index.

Human-driven Indexing: One way to build a textual index is to let experts label the database manually according to predefined concepts describing objects ("airplane flying"), scene types ("cityscape") and activities ("person playing soccer") appearing in the videos. This is done by trained experts providing *professional annotations* according to given vocabularies [AQ08]. This approach is common in large broadcasting archives or media companies. A recently prominent form of volunteer-based labeling is *social tagging*, the tagging of content in online user communities. While these approaches aim to provide searchable annotations (or tags) [GH06], they are prone to spam [KEG+08], include numbers, misspellings [CBP09], are subjective or non-relevant [UBB10], and often incomplete [BJC+13]. Another approach for label acquisition is *crowdsourcing* provided by services like Amazon Mechanical Turk or CrowdFlower [5]. Here, the labeling process is defined as a micro task which is paid and executed by a vast amount of workers from all over the world. While being considered as a valid alternative to trained experts the identification of reliable workers is important since high-volume but low-quality workers might bias overall judgments [SOJN08]. A further example for acquiring annotations are *games with a purpose* [VA06]. This approach wraps the labeling task into a game which is played for pleasure and creating labels as a side effect of playing the game. While being very successful [VABHL03] some tasks that require domain knowledge or prior training are not suitable for this type of labeling approach [VAD08].

[5] `www.mturk.com/mturk/`, `www.crowdflower.com`

In conclusion, to build an index, human-driven labeling is often impracticable due to the large amount of video data being created [SW05, Ulg09].

Machine-driven Indexing: Another way to build a textual index is to derive descriptive labels automatically by analyzing the available meta-data [WCGH99, SW05]. This can be done either by using filenames or surrounding text close to the image or video [CSBB97]. These approaches are usually used by traditional search engines. Unfortunately, in today's mobile device driven world such meta-data is often not available or not useful. For example there is no surrounded text in a personalized media archive, and given that filenames are often auto-generated during recording e.g. on a mobile phone, they lead to non-descriptive alpha-numeric strings such as e.g. *IMG_1126.avi*. Another approach towards automatic label acquisition is the expansion of social tags for indexing purpose. This can be achieved either by the expansion of available tags and the co-occurence of other tags in folksonomies as found on Flickr [SvZ08, HBU12] or by the utilization of social network structures i.e. to find descriptive tags based on the ones the uploader's friends are using [SDLW10]. Unfortunately this approach is only useful to a particular extent i.e. if tags are already associated with an image or video. They fail if users do not tag their content. An additional way to index videos is closed captioning. Closed captions are often provided for professional TV content such as news broadcasts and can be used for a linguistic analysis of video content [DZS+02]. However, although much of the content broadcast contains closed captions, personal content is often unscripted, thus not providing any closed captions.

Obviously, much of a video's information is captured by the video stream itself. Therefore great research effort is spent on *content-based* methods analyzing the audio-visual signal of a video. Besides the analysis of the audio stream [USBS12, CCC+11], automatic speech recognition (ASR) was successfully used for video database indexing [dJGHN99, HOdJ07] and is considered as a promising approach for domains such as news broadcast, political speeches or interviews. However, with the progress made in content-based image retrieval [RHC99, SWSJ00, DJLW08, DKN08], the analysis of the visual signal of a video stream became one of the very promising general purpose approaches for machine-driven indexing when no meta-data is available. With this in mind, this thesis focuses on the analysis of visual content.

1.2 Visual Learning of Semantics

A challenge of today's research endeavors at the intersection of information retrieval, computer vision, and machine learning is to answer the question if we can build machines capable of learning to perceive visual content as humans do? Once able to build such machines, we could automatically index large amounts of video content and finally provide fine grained access to unexploited video repositories.

Unfortunately, as naturally as humans are capable of perceiving their surroundings visually, just as challenging this undertaking is for machines. In the literature this is known as the *semantic gap* [SWSJ00], which describes:

> *"...the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation."*

An adaptation of this definition can also be found in [SW09], where the authors speak about the lack of correspondence between low-level features (i.e. raw pixel values) that machines can extract from videos

Figure 1.1: An overview of several visual recognition tasks. We can see a video sequence of a person riding a bicycle and performing a bike trick. The different visual recognition tasks are marked as following: **Green:** object detection, **Blue:** semantic concept detection, **Red:** multimedia event detection, **Light Grey:** affective or emotional classification, **Dark Grey**: visual sentiment analysis.

and the high-level conceptual interpretation a human associates with perceived visual content in a given situation. In this thesis, whenever a reference to the *semantic gap* is given, the original definition from Smeulders et al. [SWSJ00] is meant. This includes the objective appearance of visually present entities in images or videos such as objects ("car", "bike"), locations or scene types ("street", "beach"), as well as general activities taking place ("interview", "soccer") or the performance of complex events ("attempting a bike trick"). In contrast, whenever in this thesis a remark to a more subjective interpretation of visual content is given such as being related to affect ("romantic", "funny"), emotion ("sad", "happy"), or sentiment ("positive", "neutral", "negative"), a reference to *affective gap* [MH10] is made, which describes the mismatch between low-level features and the feelings or opinion being reflected by the image or video.

Over the last years several visual recognition tasks were defined, all aiming to achieve capabilities comparable to human perception with the goal to auto-extract different semantic and affective aspects from video content. An overview of the broad spectrum of these different visual recognition tasks can be found in Figure 1.1. The figure shows a video sequence highlighting a person riding a bicycle during daytime in an urban environment. In the middle of the video this person attempts to perform a jumping trick with his bicycle, which at first is successful but later in the video clip fails, causing an accident.Video sequences like this could be found in millions of videos on platforms such as YouTube. However, even in such a casual user-generated video clip the visual recognition task such as object detection [EVGW+10] (green frames), semantic concept detection [SOK06, SOK09] (blue captions), multimedia event detection [CCC+11, MGvdSS13a] (red frames), affective or emotional categorization [YvGR+08] (light grey bar), or visual sentiment analysis [BJC+13] (dark grey bar) can be performed. In the first couple of frames we see the object "car" and "bike" highlighted by a green frame. The second to forth frame – wrapped by a red frame – show the multimedia event "attempting a bike trick", whereas the blue captions at the bottom of the frames describe semantic concepts such as "person riding a bike", "scenes showing a street", or "outdoor scenes". The top of the figures illustrates in light grey the affective emotional labels e.g. "amazing" and in dark grey the changing sentiment from "positive" to "negative" towards the end of the video clip. Both grey bars are obviously related to the successful and failing bike trick.

It can be argued that a particular visual recognition task is a sub-task of another one or can be used as an intermediate step to solve more complex visual recognition tasks. For example, Li [LSFFX10] proposed to use a bank of multiple object detectors for scene classification with semantic concepts. In the same manner, Mazloom suggested the selection of the most informative semantic concepts for subsequent

multimedia event detection [MGvdSS13a]. This thesis will propose a similar arrangement: while focusing on concept detection, it will also utilize a detector bank to solve another visual recognition task, the task of recognizing sentiment in visual content.

Concept Detection

The task to detect generic concepts in visual content is commonly known as concept detection [NS04a, SWvG+06a, JYCN08]. Given an input image or video clip, concept detection systems use statistical learning to infer the presence of target concepts by calculating their probabilities for appearance from low-level features of the given content. Please note that throughout the research community this task has been also referred to as *image and video annotation* [FML04, WHS+06], *high-level feature extraction* [SOK06, SOK09], *semantic indexing* [SMH04, OAM+12], or *automatic tagging* [DJLW07, USKB10].

Although far from the accuracy of human annotations [YH08b, Ulg09], concept detection is of practical use as it helps to describe and understand visual content. One of the most prominent tasks it is aiming to solve is image and video search. For this purpose concept detection is applied to map visual content to a set of set of generic target concepts. These identified concepts are then used as a foundation for textual search in image and video databases [SOK09, SWdR+08]. Further applications of concept detection are content-management of TV archives [HSdRS12], content-based recommendation systems [YMH+07, RMZL12], or the support of content-based or context-sensitive advertising [MHL08, UKB12, BL12]. Finally, concept detection can be utilized for the filtering of offensive content such as pornography or violence [DPN08, JUB09, USBS12] or the suggestion of keywords for uploaded videos [USKB10, TAP+10, HBU12]. The practical applications of concept detection are of great value to cope with the massive amounts of today's growth rate of image and video content.

Usually, the set of generic concepts – or *concept vocabulary* – covered by concept detection systems consists of a broad spectrum of entities including objects ("chair", "telephone"), locations and scene types ("desert", "cityscape"), or activities taking place ("interview","people singing"). This requires concept detection systems to cover hundreds or even thousands of target concepts moving way beyond the *narrow* categorization of content [SWSJ00] with its small intra-class and large inter-class variability [SW09]. For example, even for the limited domain of TV news broadcast Hauptmann et al. [HYL07] argued that: *(i)* with a moderate detector performance, *(ii)* retrieval systems demand concept vocabularies of several thousands of detectors to reach the retrieval quality of a web search engine. Such a defined level of search accuracy can be considered sufficient for the general user.

These are two very challenging conditions. Condition *(i)* requires the construction of robust detectors dealing with well-studied issues known from computer vision such as illumination changes, occlusion, or clutter. Furthermore, it also requires generic systems to deal with different domains, ranging from TV news broadcast [NKK+05] over consumer video [JYC+11], television [AQ07b, SOU09] to web video [OAM+11, Ulg09, TAP+10]. More importantly, condition *(ii)* requires to scale up vocabulary sizes significantly. This however is considered as a major problem of concept detection [Ulg09] as it demands labeled training datasets serving as the foundation for supervised machine learning, the underlying technology of current concept detection systems [SW09]. In practice, this means that for every semantic concept to be included in the concept vocabulary we require up to hundreds of labeled samples to train a corresponding detector properly. So far, ground truth training samples have been acquired manually, i.e. a human operator labels videos or video shots with respect to concept presence. Thereby, concepts are

well defined according to a given concept vocabulary [NST+06]. This time-consuming and cost-intensive effort [AQ08, SWvG+06a, YH08a] indeed leads to high quality training material, but suffers significantly from a *scalability problem*. A second problem restricting the full potential of concept detection is its tendency to overfit to small manually required training datasets with the results of poorly generalized detectors [YH08b]. Moreover, the current setup of concept detection systems makes it infeasible to react to changing demands of users' information needs such as the timely visual detection of trending topics of e.g. sport events such as "Olympics 2012", incidents such as the "Costa Concordia" accident, or product releases such as the new "iPhone". Finally, while approaches exists to infer affect or emotion in visual content [JDF+11, WJH+12] there is a lack of methods for sentiment prediction from visual content. This kind of automatic assessment – however – would lead to a more comprehensive description of visual content in the context of online social interaction, where people express their opinion and emotions on a regular base. From the above mentioned shortcomings and drawbacks in concept detection I derive the following goals and contributions of this thesis.

1.3 Goal and Outline of this Thesis

This thesis presents strategies to address the scalability problem and its subsequent negative effects on concept detection as outlines in the previous section. The work aims to provide scalable concept learning by the reduction of manual annotation effort in using alternative training sources such as web video. This allows to synchronize concept detection with real world events matching users' information need, and enables systems to go beyond the detection of semantic concept offering sentiment analysis on visual content for opinion mining. To this end, the focus of this thesis is on the visual learning of semantic concepts with the attention towards social media driven information sources coining this combination: *visual learning of socio-video semantics*.

To achieve the above objectives, the presented strategies cover various aspects of visual concept detection systems. By aligning concept detection to users' information needs, the notion of fixed concept vocabularies has to be re-thought and a closer connection to real-world events must be established. This is done by mining social media sources for popular topics which are either mapped to an already available concept vocabulary or as a detector directly trained on demand. Starting from the underlying need for labeled training data, web video as an alternative training source is utilized. This novel information source was first exploited for semantic concept detection by Ulges [Ulg09], where user-generated tags were utilized as concept labels for visual learning. The advantages of such cost free labels acquisition – however – are limited by cumbersome query construction practice required for training data retrieval from online platforms such as YouTube and their weakly labeled nature. User-generated tags have to be considered as pseudo labels for classifier training and therefore need additional treatment or refinement. Finally, being able to provide detectors covering what people are talking about i.e. trending topics, concept detection is extended to answer how people feel about particular topics. This is accomplished by a novel enhancement at the very end of the processing pipeline: the utilization of a large-scale visual sentiment ontology represented by a detection bank of adjective noun pair concepts for sentiment analysis. Concluding, the following three contributions are presented for the visual learning of socio-video semantics:

1. enabling concept detection to adapt their concept vocabularies dynamically to user interest by on-the-fly detector training in a scalable on-demand setup.

2. detector training on web video by automatically retrieving training data and effectively handling web video's pseudo labels.

3. extending concept detection to adjective-noun-pairs enabling the prediction of sentiment being reflected in visual content.

Each strategy will be covered in a separate chapter of this thesis and is outlined in one of the following subsections.

1.3.1 Dynamic Vocabularies by Trending Topics Discovery

The first contribution of this thesis presents a novel approach towards forming dynamic vocabularies for video concept detection. The key idea is to automatically expand concept vocabularies with *trending topics* that are mined automatically on other media like Google, Wikipedia or Twitter. To achieve this, trends from different media channels are first clustered and then aggregated to form daily trending topics. An important condition to construct concept vocabularies dynamically is to predict the most popular trending topics for detector training. This is done by forecasting the life cycle of trending topics at the very moment they emerge. The presented fully automated approach is based on a nearest neighbor forecasting technique, exploiting the assumption that semantically similar topics exhibit similar behavior. Being able to identify such high-impact trending topics, this chapter evaluates several visual learning strategies for extending concept detection to auto-detect these topics in new videos, either by linking them to a static concept vocabulary, by a visual learning of trends on-the-fly, or by an expansion of the vocabulary.

Following, this work presents the first comprehensive study of various trending topics characteristics across three major online and social media streams, covering thousands of trending topics during an observation period of an entire year. Results from this study show that a typical trending topic "lives" up to 14 days with an average of 5 days. Surprisingly, the analysis indicates that Wikipedia as a media channel is as quick as Twitter when it comes to the first appearance of a trending topic. Furthermore, in real-world experiments, it is shown that on a large-scale dataset of Wikipedia page view statistics the presented forecasting method performs about $9 - 48$k views closer to the actual viewing statistics compared to baseline methods, and achieves a mean average percentage error of 45-19% for time periods of up to 14 days. This demonstrates the capability to forecast the impact of trending topics for evolving vocabularies in concept detection. Finally, in experiments on 6,800 YouTube clips and the top 23 target trends from the first half-year it is shown that a direct visual classification of trends (by a "live" learning on trend videos) outperforms an inference from static vocabularies, and that further improvements can be achieved by a combination of both approaches.

In addition, this chapter presents a concept detection system named *lookapp*, which provides real-time trending topic mining and on-demand state-of-the-art detector training. This system is built upon third-party cloud computing services (Google AppEngine and PiCloud), which allow to parallelize the construction (i.e. features extraction and classifiers training) of detectors and extend concept detection on-the-fly with new semantic concepts.

1.3.2 Training Data Retrieval and Active Relevance Filtering

A difficult challenge in concept detection based on web-video is to retrieve proper visual training content from web platforms like YouTube or Vimeo. Prior download of video content for concept learning a query has to be constructed and send to the platform to retrieve a list of matching video presumably showing the concept. As such platforms usually offer API access to their databases, the underlying query construction can be arbitrarily complex demanding a careful query construction. This chapter presents an approach which offers an automatic *concept-to-query mapping* for training data acquisition from YouTube, the largest video platform available. Queries are automatically constructed by a keyword selection and a category assignment using ImageNet [DDS+09] and Google Sets as external information sources. Results demonstrate that the proposed method reaches retrieval results comparable to queries constructed by humans, thus providing 76% more relevant content for detector training than using only concept names as retrieval queries would do.

Despite these improvements, and because web-video tags are user-generated, they can only serve as weak indicators of concept presence [Ulg09]. Such *pseudo* labeled web video contains lots of non-relevant content. So far, there are two general strategies to overcome this *label noise problem*: (1) a manual refinement supported by *active learning* sample selection [AQ08], (2) an automatic refinement using *relevance filtering* [USKB08b]. This thesis also presents a highly efficient approach combining these two strategies in an interleaved setup: manually refined samples are directly used to improve relevance filtering, which again provides a good basis for the next active learning sample selection. Results demonstrate that the proposed combination – called *active relevance filtering* – outperforms both a purely automatic filtering and a manual one based on active learning. For example, by using 50 manual labels per concept, an improvement of MAP 5% over an automatic filtering is achieved, and 6% over active learning. By annotating only 25% of pseudo positive samples in the training set, a performance comparable to training with expert annotated ground truth is reached.

1.3.3 Adjective Noun Pairs for Visual Sentiment Analysis

As the third contribution of this thesis the challenge of sentiment analysis from visual content is tackled. In contrast to existing methods which infer sentiment or emotion directly from low-level features [LFXH12, JWW+12], this work proposes a novel approach based on understanding the semantics of images. This is rendered possible by introducing a large-scale ontology of 3,000 Adjective Noun Pairs (ANP). This Visual Sentiment Ontology (VSO) is based on psychological theory [Plu80] and the proposed construction method is fully data-driven, i.e. it automatically mines online sources such as Flickr and YouTube for sentiment words, which serve as the building elements for ANPs discovery of the final VSO. This chapter also presents SentiBank, a novel mid-level representation framework, which is built upon the VSO and encodes concept presence of 1,200 ANPs from visual content. This bank of concept detectors allows the differentiation between visual concepts such as “cute dog” and “dangerous dog” and therefore allow a unique understanding of more complex labels such as sentiment. In addition, this mid-level representation of visual content can be utilized for the filtering of explicit content such as pornography or child sexual abuse (CSA) material in a way that it simultaneously provides an explanation for its detection – a system requirement demanded by law-enforcement units.

In experiments on sentiment analysis with real-world Twitter data covering 2,000 image tweets, the

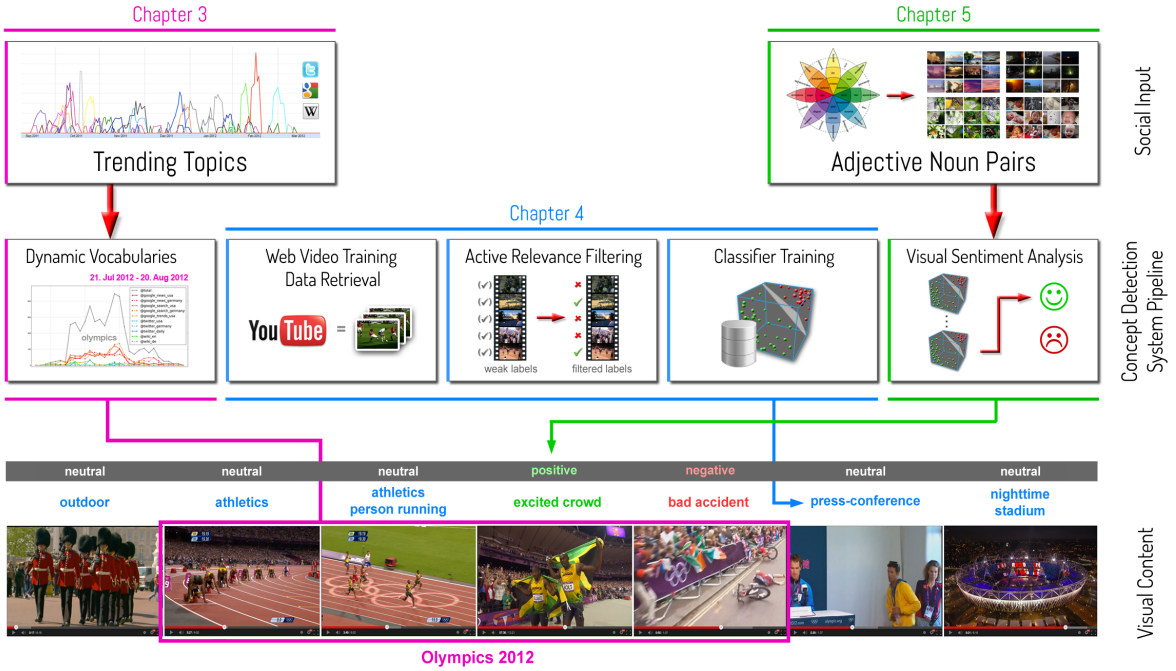


Figure 1.2: An illustration of the proposed concept detection approach utilizing social input from trending topic detection and a visual sentiment ontology. To automatically detect a trending topic (pink) in unknown visual content, the proposed framework is able to take a discovered trending topic dynamically into its concept vocabulary, retrieve automatically web video and filter non-relevant content for its detector training (blue). Furthermore, the framework is capable of analyzing visual content for the sentiment it conveys (green, red).

proposed mid-level representation demonstrates an improved prediction accuracy by 13% (absolute gain) with a joint visual-text approach over state-of-the-art text only methods. In experiments on real-world pornographic content and CSA content, the proposed approach outperformed all porn detection baselines and contributed significantly to differentiate pornographic content from CSA content (a very challenging setup, where traditional porn detection approaches lack in accuracy) Additionally, the compilation of detected ANPs allows to provide unique insights into pornographic content and CSA content. In summary, the presented visual sentiment analysis effort - being the first of its kind - creates a large publicly available resource consisting of a concept ontology, a detector library, and the training/testing benchmark for visual sentiment analysis.

1.4 Presented Framework

Concluding, the aforementioned contributions present a novel approach for an efficient visual learning of socio-video semantics from the Web. The outlined framework is illustrated in Figure 1.2: Trending Topics are mined from several social media streams. A discovered trending topic such as “Olympics 2012” is identified and added dynamically to the concept detection vocabulary. This triggers the concept detection system to automatically retrieve web video content and to filter non-relevant material for training. To detect this trending topic, the resulting detector can either be used stand-alone (pink) or be employed in combination with an already available concept detection vocabulary (blue) by a mapping

of semantic concepts such as “athletics” to the target topic “Olympics”. Moreover, in interplay with a large-scale visual sentiment ontology of adjective noun pairs such as “excited crowd” or “bad accident” the proposed framework performs sentiment analysis on the visual content of the video clip (green, red). Finally, given a video stream to be processed, the system can be used to annotate visual content on different levels of target labels such as semantic concept, trending topic, and sentiment label.

Chapter 2

An Overview of Concept Detection in Video Content

Nowadays, most video search technology, be it for large-scale online video platforms such as YouTube or video archives of television broadcasters, rely on human-driven indexing i.e. manually generated descriptions, annotations or tags. As seen in Section 1.1 this type of indexing is prone to spam, misspellings, subjectivity, and is incomplete or non-relevant and therefore not practicable for many applications. Even if done by professional annotators for archiving purposes, this approach does not scale with the immense amount of audio-visual content currently being produced [WCGH99, Sme07, SW05]. A solution to this problem is concept detection, a machine indexing mechanism, which provides access to video content by analyzing the audio-visual video stream for the presence of semantic concepts such as objects ("chair", "telephone"), locations and scene types ("desert", "cityscape"), or activities taking place ("interview","people singing"). Throughout recent years of intensive research the academic community referred to this challenging task also as *image and video annotation* [FML04, WHS+06], *high-level feature extraction* [SOK06, SOK09], *semantic indexing* [SMH04, OAM+12], or *automatic tagging* [DJLW07, USKB10]. In this thesis the widely established term "*concept detection*" [SW09] will be adopted.

This chapter provides an overview of concept detection research with a focus on video content. Approaches which are specific to the contributions of this thesis (as outlined in Section 1.3) are described later in their corresponding chapters. This chapter starts with the outline of the problem statement and the definition of concept detection (Section 2.1). It further lists its most important application areas (Section 2.2) and provides an introduction to the structure of video material (Section 2.3). Then, the chapter continues with an overview of concept detection in the context of employed architectural frameworks (Section 2.4) and frequently used approaches and methods (Section 2.5). To this end, it introduces common means fÃijr system evaluation (Section 2.6) and closes with an examination of label acquisition in Section 2.7.

(a) (b) (c) (d)

Figure 2.1: Example keyframes of YouTube videos displaying the presence of different semantic concepts. As seen, **(a)** different textual descriptions such as "person riding a bike" or "bike stunts" might match the same visual content, **(b)** descriptions are subjective or require prior knowledge to recognize e.g. a "bride", **(c)** require exact specifications in form of restrictions to physical objects only excluding animations of objects, or **(d)** may display the semantic concept (e.g. "shots of an airplane flying") from an unusual point of view.

2.1 Problem Statement

Concept Detection aims to analyze the audio-visual content of video to automatically infer the presence of semantic concepts. To make this compact description more formal, the notion of a "semantic concept" and "concept presence" has to be provided in more detail. As outlined in the introduction of this chapter, a semantic concept can be anything from the broad range of objects, locations, scene types, or activities taking place. Snoek and Worring define a semantic concept as: *"... an objective linguistic description of an observable entity"* [SW09]. Although this definition does capture the idea of a concept itself and its presence in video content, a more specific notation is desirable. Consider the following five examples of concepts definitions as illustrated in Figure 2.1:

c_1 := "a person riding a bike"

c_2 := "bike-stunts"

c_3 := "attempting a bike trick"

c_4 := "bride: a woman on her wedding day"

c_5 := "shots of an airplane flying"

It can be seen that the three concept definitions c_1, c_2, c_3 match the visual content in the keyframes of Figure 2.1 (a). Interestingly, they all come from different backgrounds and serve different purposes: c_1 describes a NIST TRECVID concept of the *Semantic Indexing Task (SIN)* [OAF+10], c_2 is retrieved from the actual YouTube video, where the uploader of the video decided to assign the tag "bike-stunts" to his video clip, and c_3 can be found as a description of one of the events in the *Multimedia Event Detection* task [OAM+13]. Although all three concept descriptions come from different recognition tasks or are created with different intentions (intrinsic as given by a YouTube user or explicit as defined by the NIST TRECVID annotation protocol) they might match the same visual content i.e. agree to concept presence in Figure 2.1 (a). Furthermore, the initial definition of a concept by Snoek and Worring [SW09] emphasizes the importance of objectiveness in a concept description. However, looking at Figure 2.1 (b) and c_4, which displays a keyframe from the wedding of Prince William and Kate Middleton, depicts a "bride, a woman on her wedding day", a semantic concept from LSCOM (ID: 132) [NST+06]. This

might look like an objective match of concept presence for c_4 and a person who grew up in a western culture, but it might lead to disagreement for somebody with a different cultural background, rendering this concept definition to be subjective rather than objective. As illustrated, even for such apparently well-defined concepts, concept presence might be judged differently by different people.

Nevertheless, this difficulty in providing a concept definition can be eased by establishing further rules or protocols guiding the understanding of concept presence. The NIST TRECVID benchmark is one of the leading instances doing so for video content[1]. For its SIN task the definitions of concept presence are further enriched by restrictions such as: *"physical objects representing the target, such as photos, paintings, models, or toy versions of the concept, will not be grounds for judging the concept to be true"*. Such a restriction is illustrated in Figure 2.1 (c). The top keyframes shows an Airbus 380 during its test flight obviously displaying the concept c_5, an airplane flying. However, considering the bottom keyframe of an animated airplane flying – according to the restriction of the NIST TRECVID SIN task – this keyframe does not contain the semantic concept of an airplane flying. Similarly, for NIST TRECVID's MED task, the organizers decided to provide event kits describing the event concept by additional information such as its event name, definition, explication (textual exposition of the terms and concepts), evidential descriptions, and illustrative video examples [OAF+10].

Although such additional restrictions can help to convey what is meant by the semantic concept, they unfortunately can not solve the problem of ambiguity entirely. For example, Figure 2.1 (d) depicts some keyframes for concept c_5 demonstrating ambiguity of a concept description. The keyframes show the inside of an airplane flying as tagged by some YouTube users. In such a case the description of an airplane flying is correct. It even confirms the previously mentioned SIN protocol for concept judgments – however – it displays the concept from a very different point of view, which increases the difficulty to recognize the concept, even if user assessment of the concept was correct.

Altogether, this has the following consequences; First, we recognize that different concept descriptions can be correctly depicting the same concept and the same textual concept description can lead to a high variability of visual content. This is especially true for web video as found on platforms such as YouTube. Furthermore, it is important to note that even with a clear textual description of a semantic concept the corresponding visual content representing the concept as seen might not be captured entirely or be ambiguous. Still – to provide a canonical definition of semantic concepts, this thesis adopts the definitions from Ulges [Ulg09], which considers *relevance* as the defining criteria for semantic concepts taking its high variability of real-word user generated content on YouTube into account [Ulg09, NMP10, TAP+10].

Definition: "Concept Detection" Concept detection is the task of inferring if a semantic concept c is present in a video X i.e. if a concept c is visible in a video X. A concept detection system is built upon a set of target concepts $Voc := \{c_1, ...c_n\}$ and analyses incoming video clips, which can be of different structural granularity (please refer to Section 2.3 for more details). The goal of concept detection is to estimate *scores* $\phi_{c_1}, ..., \phi_{c_n}$ indicating whether a concept $c_i \in Voc$ appears in X. This output may also be interpreted as a probability of concept presence for c_i (often after transformation by an appropriate monotonic function). But for many retrieval applications it is sufficient to directly rank videos by sorting them according to their scores. The described multi-class scores estimation is usually divided into multiple binary classification problems, i.e.the score ϕ_{c_i} for each concept $c_i \in Voc$

[1] `http://trecvid.nist.gov`

is calculated independently and correlations between concepts are considered in subsequent processing steps.

The most prominent method to model the mapping ϕ_c is by using a statistical classification algorithm (e.g., an Artificial Neural Network or a Support Vector Machine). Such supervised machine learning approaches require a training step prior to their application, to estimate model parameter for the final classifier $f_c := X \rightarrow \phi_c$. For this purpose, concept detection systems are usually separated into an offline phase for training and an online phase for application. For classifier training, the system requires positive and negative samples, i.e. training videos with *labels* denoting the presence or absence of the concept c, $\mathcal{D}_c := \{(x_i, y_i) \mid x_1, \ldots, x_n, y_i \in \{-1, 1\}\}$. These labels have to be acquired up-front classifier training either by experts providing *annotations* according to a controlled vocabulary or by YouTube or Flickr users providing *tags*. This thesis adopts this differentiation between the notations of *labels*, *annotations*, and *tags*.

2.2 Applications

Current concept detection systems apply machine learning, allowing to scale up vocabulary of target concepts if labeled training samples are available. Similarly to information retrieval, which is concerned with the representation, storage, organization, and access of information items [BYRN+99, SM86], concept detection focuses on the analysis of video material to bridge the semantic gap with the goal to provide descriptions of video content. And although the performance of such systems with respect to quantity and quality is far from optimal [HYL07, YH08b], recent improvements in content analysis are promising [SS10] and the practical benefit of such systems would allow to advance in the following areas:

Video Search As introduced in Chapter 1, one of the major goals of concept detection is to render textual search on video possible. Such applications are usually based on a fixed vocabulary of a target concept and a query processing engine utilizing the underlying index of detected semantic concepts [CH05, SWWdR08, SWdR+08]. One of the most prominent research efforts in this area is the TRECVID Search task [OAR+09], which aims to provide search and browsing tools for human analysts, who are looking for segments of video clips containing semantic concepts, which might be peripheral or accidental to the original subject of the video. Such a query processing can be either realized by the use of a vector space model to match a query against the semantic description of a concept [NZKC06, SWvG+06b], the restriction to query classes [YYH04, ZSC+06], local context analysis [YH06], or the use of external sources [Fel98, KNC05]. For an evaluation of video search using concept detection please refer to [NHT+07]. Although such a mapping of textual queries to visual content strongly depends on the quality of concept detection systems, which is far from careful manual annotations [YH08b], it can be considered as a key building block of modern content-based video retrieval systems [SW09].

Video Tagging With the emergence of online video sharing platforms the amount of video content being uploaded has increased rapidly. On platform like YouTube, every uploader of a video clip is asked to tag his content i.e. define annotations for his video. Concept detection can help in recommending tags, which semi-automatically can be selected by the user [ATY09, TAP+10]. This process of selecting a subset of possible tags, although not effortless, is more convenient for the user than defining his own

tags. Contrary to recommending tags during the upload of video content, the concept detection of video material can be performed retrospectively on the entire video database to auto-tag or predict tags for video clips providing a basic indexing [WHS+06, CEJ+07, NMP10, YT11].

Video Recommendation While video search is a very active task for a user, video recommendation is a passive mechanism of video consumption. Similar to tag recommendation, users find the binary decision (decide whether to watch a recommended video clip or not) more convenient than the sometimes exhausting work of formulating the right search query. Hence, video recommendation plays an important role in the context of content discovery unknown to the user, which play an important role on video-on-demand platforms and online video sharing platforms [DLL+10]. The visual analysis of video content by concept detection is considered as one of the alternatives available to cold start (i.e. no user history) video recommendation. Such systems can therefore be used for either the recommendation of video a digest [YMH+07] or the personalized delivery of video content [LFKS09].

Content-based Advertising in Video One further application of concept detection is realized in the context of targeted advertising. As video distribution is a costly venture [Sil] it requires more sophisticated monetization channels than traditional TV broadcasting. One promising instrument in this regard is the semantic linkage of advertising with the content of video clips [SSW07, BLS01, MHL09] or images [WYZ+09, MHL08]. Similar in motivation is the prediction of demographic groups (gender, age) for advertising by the identification of semantic concepts present in video clips [UBK13]. Here, concept detection plays a crucial role in situations where a video clip is freshly uploaded (i.e. viewer statistics are unknown) and only little information is given by the uploader.

Video Archiving The current rapid growth of online multimedia collections is not the only source of video material to profit from concept detection. As long as TV broadcasting exists, visual content produced is archived. This material, although digitalized, is unfortunately not made accessible due to the annotation effort associated with it. Here, concept detection is playing a crucial role in granting efficient access to such digital archives maintained by TV stations. Especially news broadcasting and documentaries with contemporary witnesses reporting about historical events are of interest because of their role to preserve our cultural heritage visually [YOU11]. With respect to this application scenario concept detection systems can be trained to cover a controlled vocabulary of target concepts, providing automatically searchable annotations [HSdRS12]. This way such archives can be made accessible for either educational or journalistic purposes.

Content Filtering The recent advances in network technology allow for seamless distribution and sharing of all types of visual content. Unfortunately, this circumstance is exploited for the unrestrained spread of offensive, harmful, and illegal video material over the Internet. The forensic detection of this material poses a difficult challenge as police forces find themselves confronted with a flood of digital content e.g. during their fight against child sexual abuse (CSA). A concept detection approach can be used to either identify specific content and therefore reduce the amount of manual investigation needed, or to filter specific content for parental control. Such approaches were already proposed to detect violence [DSDVL02, LW09], nudity [USBS12, JUB09, DPN08] or illegal pornography[US11] in visual content.

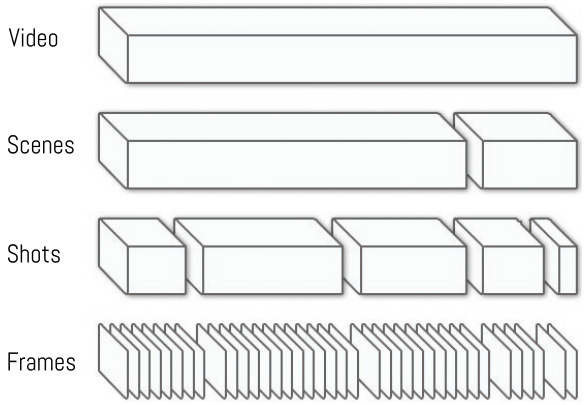


Figure 2.2: An illustration of the hierarchical structure of a video clip organized according to units of temporal granularity. The entire video clip (first level) can be organized in scenes of narrative context (second level), which itself can be split into shot defined by a single camera recording operation (third level). Each shot is comprised at the lowest level of a set of adjacent frames (fourth level).

As outlined above, the range of applications, which can be built upon concept detection is broad and tackles not only commercial interest but can also help in the context of societal hurdles. These applications render the automatic detection of semantic concepts as a rewarding research area. Having said that, before we go into detail about how state-of-the-art concept detection system look, we should look how video as a medium is structured.

2.3 Video Structure

A video clip is organized in temporal units that define a chronological story for the audience [Mar04, SS02]. Nowadays in the age of mobile phones, user-generated content and platforms like YouTube a narrative storytelling may not always be recognizable but nevertheless the fundamental hierarchical structure [NH01] as illustrated in Figure 2.2 is still a valid representation of video clips.

This structural composition usually consists of different levels of temporal granularity. On top of this hierarchy, the entire *video* clip can be seen as a global unit of the content enclosed. This first level is usually further organized in *scenes* of narrative context such as e.g. dialogue, atmospheric or transition scenes, just to mention a few prominent scenes or logical story units [HLB99]. These scenes can be split into *shots*, the basic unit of motion picture production. A shot is defined as a sequence of continuous frames that are recorded through a single camera operation. Two shots are concatenated by a *transition*, which can be either abrupt (e.g. a hard cut or a black frame) or gradual (like wipes, dissolves, fades). The last level in this hierarchy is the single *frame*. Alone in isolation a frame is nothing more than a spatial plane, rasterized into pixels, equivalent to a digital image or photo. However, exploited as a consecutive sequence of frames it enables the illusion of motion in video.

With respect to content analysis for concept detection, the lower two levels (the frame level and the shot level) are of particular interest. While the shot level provides temporal information, which can be beneficial for detection, the frame level allows for the spatial analysis of content as known from content-

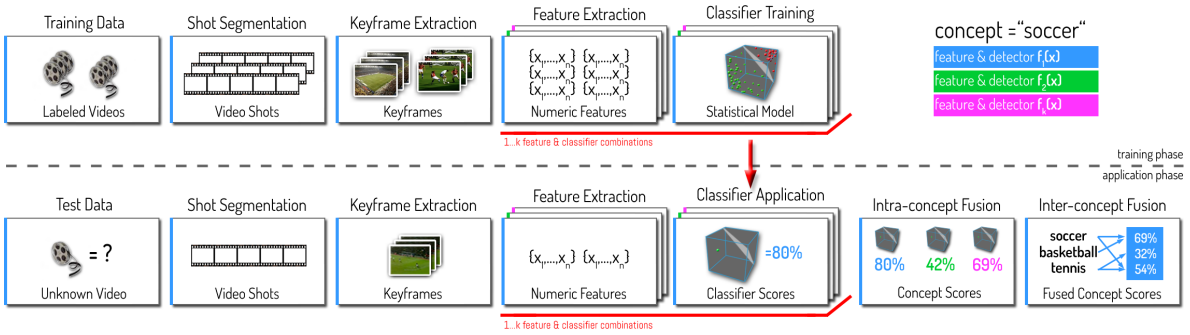


Figure 2.3: An overview of a current state-of-the-art concept detection system. For each target concept a separate training phase (top pipeline) is required to construct a statistical model which is specific for this target concept. This model will be used in the application phase, where an unknown video is tested against the previously trained classifier. Furthermore, this overview also illustrates common components of a concept detection processing pipeline.

based image retrieval (CBIR) [SWSJ00] and is therefore considered as the basic unit of analysis in many state-of-the-art concept detection systems [ABC+03, JYCN08, SW09, USKB10].

This is also driven by to the availability of concept annotations for the visual content of a video clip. Originally, annotations are given at shot level of a video clip [SOK09], However, because of the adjacency to image analysis, it is also common to work with annotations on frame level to allow for more temporal accuracy of content description [USKB08b]. In contrast to this, an exception has appeared with the rise of user-generated video on the Internet, where videos are tagged by the uploaded. Here, only global annotations for the entire video clip are provided (which is subject to this thesis and handled in Chapter 4). Further, the frame and shot level also define prominent fragments for the evaluation of concept detection system performance as seen in more detail in Section 2.6.

2.4 Concept Detection System Architecture

A major research effort in the context of concept detection is the TRECVID [Sme05] benchmark, which initially addresses concept detection in its *High-Level Feature Extraction* task [SOK09] and now in its *Semantic Indexing (SIN)* task [SOK09]. This benchmark aims to evaluate the performance of different concept detection systems on common, standardized datasets, compare results and allow an exchange of experience within the research community.

One of the most notable observations in this regard is the dominance of systems utilizing supervised machine learning as an underlying technique for the inference of concept presence in videos [CEJ+06, NJW+09, ea11]. This agreement of how concept detection systems are built nowadays is also reflected by the definition of concept detection from Section 2.1. Concluding, supervised machine learning provides a generic strategy to construct classifiers, each specifically trained to detect one particular target concept. Therefore an intrinsic property of such systems, is the separation of a preliminary *training phase*, which usually takes place offline, and an *application* or *testing phase*, which composes the online detection process [SOK09, SW09]. Such a system architecture can be seen in Figure 2.3 as described in more detail in the following.

2.4.1 Training Phase

The goal of the training phase is to create a statistical model specifically trained to detect the presence of a target concept such as e.g. "soccer scenes" in visual content. Such a training step is mandatory and must be performed for each of the target concepts $c_i \in Voc$, defined by the vocabulary of the concept detection system. Also, the construction of statistical models – often also called classifiers or detectors – is often very time consuming but nonrecurring for the lifetime of such a system, this phase is usually completed offline.

To train a single classifier f_{c_i} for the target concept c_i, one requires a labeled training dataset $\mathcal{D}_{train} \in \mathcal{D}_c$. Such a dataset is a labeled set of visual samples $(x_i, y_i) \in \mathcal{D}_{train}$ representing the target concept. Its acquisition is usually done by experts manually annotating video data [AQ08] or recently by the use of tags from web sharing platforms such as YouTube [Ulg09, USKB10, UKBB09], Flickr [LSWS12] or both [KLS13]. Both label acquisition sources have their advantages and disadvantages (see Section 1.1). While the manual process provides very accurate labels it is lacking in scalability due to the very time-consuming effort which is associated with it. Using tags as labels solves the labeling effort because of their free availability on web platforms but they are facing other challenges as discussed later in this chapter, in Section. 2.7.

A common procedure in detector training is to first segment all training videos into shots and to extract representative keyframes. Then for each keyframe, features are extracted describing its content in the form of numerical values. These features are then used for classifier training, yielding a statistical model ready for the application phase. It is common in current concept detection systems to train several feature (and classifier) combinations for the same concept if the features can provide additional clues about visual content such as with color features and texture features. As illustrated in Figure 2.3 this would technically imply to train three individual statistical models (indicated by the colors blue, green, and pink) for the same concept, but each with different features selected as input.

2.4.2 Application Phase

The goal of the application phase is to detect the presence of a target concept in unlabeled videos. To accomplish this, the output of the training phase – the statistical model, representing a target content in features space – is used to analyze an input video. Since the application of a classifier is usually less time consuming, this phase of a concept detection system is often realized online.

Similarly to the training phase the input video is segmented into shot level and keyframes are extracted. These keyframes are processed to extract the same numeric features describing their visual content as in the training phase and simultaneously each feature is then fed into its trained statistical model for the target concept. On concept level this includes all feature classifier combinations available from training. As an output, these classifiers provide detection scores, which differ from feature to feature. For example the concept "soccer scenes" will have a different detection score when using color information than when using texture information rending the detection scores as complementary information for the visual presence of the target concept. These individual scores are then fused into a single concept score (intra-concept fusion) for one single target concept. As a last step, a concept relation modeling (inter-concept fusion) is performed, which takes concept correlations into account to refine the final detection output accordingly. Such a refinement is feasible due to the correlation of supporting

concepts like “ocean scenes” and “boat / ship” and the suppressing relation of concept like “airplane flying” and “person riding a bike”.

As seen in the above description and Figure 2.3 both phases, training and application have a very similar pre-processing of video material. The conceptual difference is that whereas in the training phase, feature classifier combinations are constructed (i.e. detectors are built), in the application phase they are applied on an unknown video clip. Specifically, the post-processing with is intra-concept fusion und concept relation mapping (inter-concept fusion) is specific for the application phase allowing it to form a final concept detection score. A detailed review of each component in such a processing pipeline will be covered in the next section.

2.5 Concept Detection Pipeline

The processing pipeline of most concept detection systems [SOK09] is, as observed, closely aligned along the hierarchical structure of video as introduced in Section 2.3. Therefore an obvious conclusion is that additional processing steps for video analysis are necessary to the ones known from content analysis on images [SWSJ00, DLW05, DJLW08]. Due to the nature of video, its processing pipeline must additionally be able to handle the temporal dependency and relationship of individual frames, which creates new semantics that may not be present considering an isolated single image or frame (e.g. motion information). To this end, a concept detection procession pipeline can be described by six major components as seen in Figure 2.3 (application phase): shot segmentation, keyframe extraction, features extraction, statistical classification, intra-class fusion, and concept relation modeling (inter-class fusion) [SW09]. Next each component of such a pipeline will be described in more detail.

2.5.1 Shot Segmentation

Given the natural structure of video content as seen in Section 2.3, one of the first steps in video analysis is the temporal segmentation of a video clip into shots [ABC$^+$03, JYCN08, CEJ$^+$06, NJW$^+$09, ea11]. A shot is also one of the basic units of annotations, analysis, and evaluation in benchmarks such as TRECVID [OAM$^+$12]. This task is usually approached by *shot boundary detection*, which aims to detect shot transitions in a video stream according to sudden changes in the visual appearance of successive frames [TRE07]. An example can be seen in Figure 2.4 [BUSB08]: the figure illustrates pair differences of visual appearance represented by a feature descriptor over time. Peaks depict candidates for cuts defining the boundaries of two subsequent shots. Three types of shot boundaries are basically recognized: hard cuts, dissolves and wipes. Fade-in and fade-out are usually defined as dissolves either starting with a black screen or ending with a black screen. While hard cut detection can be reliably solved by known algorithms [Lie99, Han02, YWX$^+$07], dissolves and wipes are more difficult to detect. Approaches to detect dissolves and wipes are based either on edge change ratio and standard abbreviation of pixel intensity [Lie01] or luminance pixel values [Pet04]. Although a challenging task in the early years of video analysis, shot boundary detection is nowadays well understood and considered solved by the research community [SOD10].

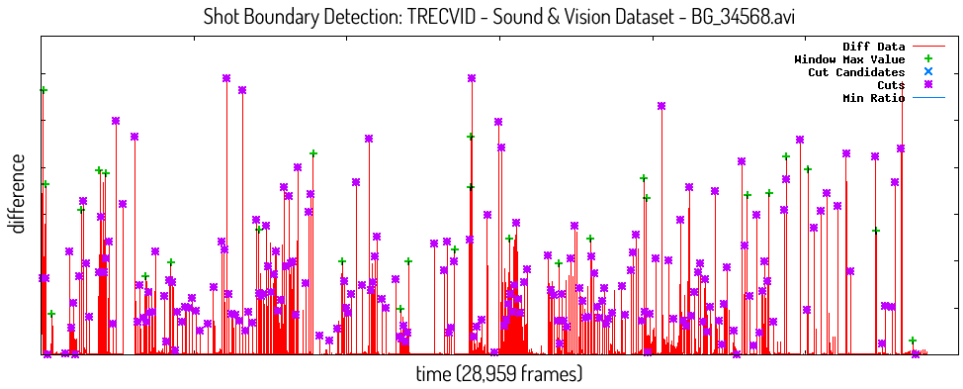


Figure 2.4: An illustration of a shot boundary detection approach, which is driven by change detection of spatial features of video frames. The peaks depict local changes on feature level and indicate a possible cut in the video stream. The example is taken from [BUSB08] and illustrates cuts detected in a TRECVID Sound & Vision sample video.

2.5.2 Keyframe Extraction

To capture the content of a video it makes no sense to analyze the spatial domain of every single video frame given that subsequent frames are very similar to each other with little information gain between them. Additionally the immense amount of frames defining a video makes it time-consuming to process every frame [ZRHM98]. A general method to handle such a huge amount of content is to extract representative *keyframes* conveying most of the content of the video. Such a reduction of a temporal video stream to a set of characteristic keyframes also has the advantage to enable the usage of known analysis techniques from image retrieval [Hau05].

Several extraction methods to extract keyframes have been investigated in the literature. One of the most straightforward ones is the selection of a single frame as the keyframe – this can be either the first, last or middle frame of a shot [O'C91]. Although very prominently used in TRECVID [SOK09], this keyframe extraction obviously loses information in longer shots as compared to the extraction of multiple keyframes per video shot [SWG$^+$05]. Another method going in the opposite direction is regular sampling along the video stream [USBS12]. However, the advantage of a dense sampling of video content remains in contrast to the large amount of keyframes being extracted. A group of methods in-between is adaptive sampling of keyframes. These methods are based on the complexity of video content and extracts keyframes either by strong content change [ADDK99, UKBB09] or unsupervised learning via clustering and the designation of cluster centers as keyframes [ZRHM98, HZ99, HM00, MRY06, USKB10].

A visualization of keyframe extraction methods can be seen in Figure 2.5. The previously mentioned groups: *single frame*, *regular sampling*, and *adaptive sampling* can be either applied on the entire video (Figure 2.5 (a)) or the segmented shots of a video (Figure 2.5 (b)). As seen, the single frame method is not recommended to be used on long shots or the entire video, whereas in the case of regular sampling the temporal segmentation into shots does not have an impact on keyframe extraction. With respect to adaptive sampling the application on video or shot level may have an impact on the way keyframes are extracted since such methods find an adequate number of representative keyframes for the given shot with respect to its visual complexity. Please note, that keyframe extraction – besides its use for content

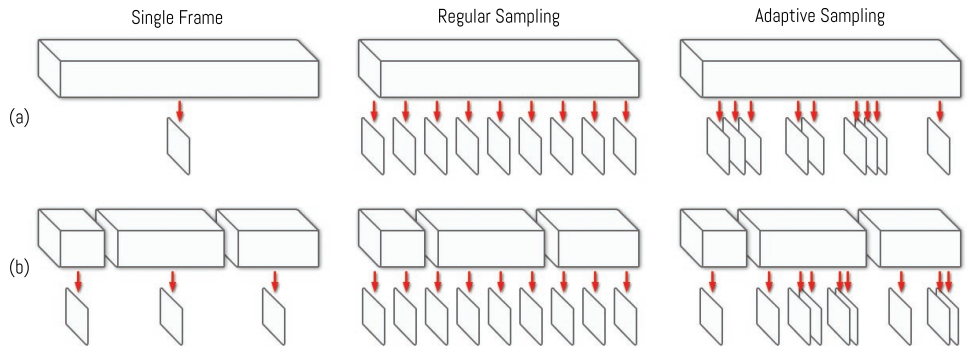


Figure 2.5: A digest of different keyframe extraction methods. In general three different groups of methods exist: *single frame*, *regular sampling*, and *adaptive sampling*, which can be either applied on the entire video clip (a) or on shot level (b).

analysis – is also of use for browsing [NH01, BSUB08] or summarization [MRY06] of video content.

2.5.3 Feature Extraction

The aim of feature extraction is to transform shots or keyframes into feature vectors $x \in \mathbb{R}^n$, which can be used as input for the subsequent classification step. This procedure results in a feature space representation of the video content. With respect to this, feature descriptors should be discriminative, to a particular point invariant, and computationally inexpensive to be extracted from the content. Two major classes of features can be distinguished in video analysis: feature descriptors based on the temporal video stream such as motion or audio, and features descriptors based on the spatial analysis of keyframes. The latter can further be separated into global or local features. In this context different descriptors have been proposed, ranging from global color, texture and shape descriptors [DKN08] to local patch-based ones like the very prominent *bag-of-visual word* representation [SZ03] with SIFT [Low04] or SURF [BTvG06] features. Especially patch-based descriptors proved to be robust and give high accuracy in several computer vision tasks [EVGW+08, JNY07, vdSGS08b].

The following provides an overview of major feature types used in concept detection systems. For more information on feature extraction please refer to the evaluations in [DKN08, SWSJ00, vdSGS08b].

Color Color perception is an important element of the human visual system. Widely used methods of color features are their global statistical distribution e.g. RGB color histogram [SC97, WLL+07] or derived from that, RGB color moments [NH01]. While color histograms have the advantage of being invariant under rotation, they represent an image globally not considering the spatial structure of color. To capture the spatial location of color, further descriptors have been presented such as e.g. the MPEG-7 defined color descriptors: Color Structure Descriptor, Scalable Color Descriptor, Color Layout Descriptor, Dominant Color Descriptor [Mar04, MOVY01]. Other layout preserving approaches, which bin or partition the frame into a grid also exist, such as the spatial pyramid representation of color histograms [vdSGS10], or combined global histograms with spatial information as done in the color correlogram descriptor [HKM+97].

Texture Another global feature descriptor group is based on properties of image textures [MM96]. Texture information is never isolated to a single pixel but rather exists within a region of pixels defined by its local neighborhood. One method to extract such neighborhood information from textures is to use different filters against the image e.g. Gabor filters [ZWIL00, GPK02], Wavelets [MM96], or spatial-frequency based ones [MOVY01]. Another idea is to define textures according to their coarseness, contrast, directionality, line-likeness, regularity and roughness [TMY78], which proved to be a robust and useful descriptor for image retrieval [DKN08]. Textural cues can also be combined with color information as presented in color invariant texture [vGV$^+$06]

Edges Edges define prominent properties of images and often occur in conjunction with image texture characteristics [MOVY01]. Such descriptors are often composed as histograms over orientation of edges [MOVY01, WLL$^+$07], which have been found in the image by edge detectors such as Canny [Can86] or Harris [HS88]. Similarly to texture, edge based descriptors can also be combined with color to render them color invariant [vGV$^+$06]. A related descriptor close to this group of edge descriptors is the histogram of oriented gradients (HOG) descriptor [DT05], which generates histograms of image gradients instead of image edges. This method is known to be highly efficient in concept detection when applied on local patches [TAP$^+$10].

Shape Shape-based feature descriptors are motivated naturally by the idea that man-made objects have typical geometric structures and shapes. Basically two groups of geometric shape descriptors exist: contour-based shape descriptors and region based shape descriptors [Bob01]. Properties which are used as descriptors are perimeter, area, compactness, contour Fourier coefficients or geometric moments [NBE$^+$93]. As motivated, such features are useful for object category recognition [JS04] but also can improve image retrieval [NBE$^+$93]. Another representative shape based feature is the GIST [OT01] descriptor. This global descriptor extracts the spatial envelope of a scene and provides a low-dimensional set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene.

Patches Similar to local edge or shape feature descriptors, a third group of local descriptors is used in computer vision research. These so-called patch-based feature descriptors are characterized by their high robustness against clutter, deformation, and partial occlusion and often come with invariance against scale, orientation, and illumination [Lin98, KB01, MCMP02, MS04]. This is achieved by the detection of prominent, salient image patches – so called *interest points* – which serve as local regions for feature extraction. Most prominent representatives of patch-based descriptors are SIFT [Low04] and SURF [BTvG06]. These features can be considered as the best-performing descriptors in several visual recognition systems [SZ03, FML04, HL04, DKN05, FFP05, MLS06] and benchmarks [EVGW$^+$08, SOK09, DDS$^+$09, CCC$^+$11]. Additional details about patch-based descriptors can be found in the following surveys [SMB00, Mik03, Rot08, vdSGS08b, CLVZ11]

Bag-of-Visual-Words Frequently used with patch-based feature descriptors, *bag-of-visual-words* representation of visual content has gained in popularity over the last decade [SZ03, FFP05, SREZ05, QMO$^+$07, ZMLS07]. This feature representation is motivated by the bag-of-words model in text analysis [Lew98]. Similar to textual documents which can be represented by counts or word occurrences, an

image or visual document is represented by counts of *visual worlds* from a visual codebook. Built upon patch-based descriptors such as SIFT, the construction of a bag-of-visual-word descriptor is done are follows: as described above SIFT represents an image as a set of interest point descriptors. This structure – however – is varied in cardinality and lacks a meaningful order. Since classification models usually require feature vectors of fixed dimension as input, a vector quantization technique is applied to partition the SIFT feature space into a large number of clusters. The clustering process generates a codebook of visual words describing different local patterns in images. The number of clusters determines the size of the codebook, which can vary from hundreds to thousands [PCI+07]. Consequently, each SIFT descriptor can be encoded by the index of the cluster to which it belongs, which automatically assigns it to the element of the codebook. Counting the represented visual word elements for each SIFT descriptor in an image leads to the final *bag* representation. Different extensions of these feature descriptors have been presented in the literature, such as hierarchical setup [LSP06] or soft assignments [PCI+08, vGVSG10, CLVZ11].

Text Text in video can be another valuable source for feature descriptors in concept detection [WCGH99]. Text can appear in videos as scene text (e.g. logos on buildings), overlay text (e.g. name of the displayed person), or closed captions. The task of transforming such text into machine readable text is usually split into text detection and optical character recognition (OCR) [LDK00]. While being being more challenging than traditional OCR [WBB11], research effort in this area is actively pursued on detection of scene text [SSD11].

Mid-level-Representations In contrast to the previous low-level features, this group of feature descriptors introduces a mid-level attribute representation of visual content [FZ07]. This representation is motivated by the observation that a classifier output can be used to recognize unseen object categories from their description in terms of attributes [LNH09, FEHF09, KBBN09].

Following this mid-level feature representations take the output of low-level feature classification as input for a subsequent learning of target concepts. Examples in this area are the discovery of visual attributes [FZ07, BBS10, LNH09, FEHF09, YJT+12, RFF12], the construction of signatures from large concept detection vocabularies [HvdSS13, MHS13, TSF10] or the compilation of classifier banks such as ObjectBank [LSFFX10], DetectionBank [ASD12], or ConceptBank [MGvdSS13b]. This kind of feature representation became a promising research direction in recent years. It builds upon the vast amount of available training data and computational resources to construct large-scale collections of classifiers. In particular this type of feature proved to be successful in the detection of complex constructs like multimedia events [SvdSF+13, BCC+13].

Motion The analysis of temporal relations in video enables the acquisition of information, that elsewhere would have been lost. Different than the previously described keyframe based features, this type of feature introduces the concept of motion as extracted from video shots. Motion, i.e. the change of a location in time, translates in digital video into the spatial location change of pixel blocks over consecutive frames [ACAB99]. Unfortunately such an observation of motion provides no real differentiation between camera motion and object motion or multiple object motions [BA96]. Motion features are usually extracted as 2-dimensional motion vectors in the image plane either by the tracking of spare but salient features [TK91], an optical flow estimation [BB96], spatio-temporal pixel regions [DD03], or

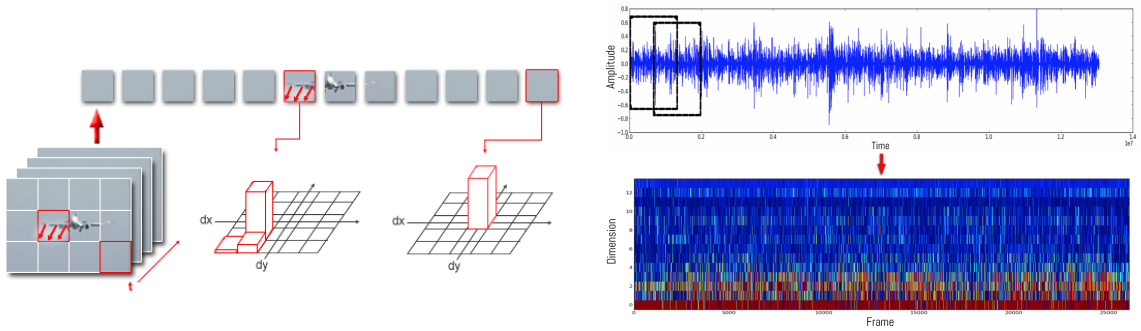


Figure 2.6: **Left:** Illustration of a motion histogram descriptor. A motion vector field extracted from consecutive frames is partitioned into 12 blocks. For each block the dx and dy contribution of all motion vectors is accumulated into a histogram [USKB10]. **Right:** MFCC audio descriptor extraction from an audio signal. A sliding window is moved over the signal and for each overlapping frame its Mel Frequency is computed and concatenated over time [Bad11].

texture pattern extracted from the motion vector field [MZ03]. Often they can be directly extracted from the compressed video stream [ACAB99, USKB10]. Once a motion vector field is extracted it can be used to build motion histograms as shown in Figure 2.6 (left) serving as input for concept classification [HN07, USKB10]

Audio An orthogonal modality to the visual content in video is the analysis of the audio stream in video clips. To this end, methods such as automatic speech recognition (ASR) or background noise analysis can be employed to provide insights about the content of a given video. In the case of ASR analysis, the goal is to analyze the spoken word and deliver it as text for search and retrieval. While for some types of video content e.g. for news broadcasts, spoken words do have a strong alignment with the displayed content [CMC05], this alignment is not always guaranteed. For example it may be possible that two people are speaking with each other about "mountains" but no mountain is visually present in the video at the moment the concept was mentioned. Nevertheless, in the past it was successfully used for retrieval in audiovisual archives [Sme07, HOdJ07, HSdRS12]. Background noise analysis can also be utilized as clues for concepts in video clips. For example, the barking of a dog is a strong indicator of the visual presence of a dog in a video clip. Such analysis often combines the prominent bag-of-word approach with Mel Frequency Cepstral Coefficients (MFCC) audio features (Figure 2.6 (right)) for either the detection of multimedia events [CCC+11, RLFF13] or for filtering of specific video content [LW09, USBS12].

2.5.4 Statistical Classification by Supervised Learning

As shown in the previous section different approaches exist to extract global and local features from a keyframe or a video shot. These feature extraction approaches extract n-dimensional vectors directly from the raw pixel or audio information and provide as a result an n-dimensional *feature space* representation of the corresponding keyframe or shot. Given such a feature space representation for a sample $x \in \mathbb{R}^n$, the goal of statistical classification is to infer concept presence or absence by the estimation of a numerical score ϕ_c of a target concept c. These scores can be either directly taken to rank classification results or

in a probabilistic setting, can be interpreted as a posterior of concept presence. In concept detection, ϕ_c is usually modeled as a binary classification problem. As common in supervised machine learning, model learning is performed on a set of training samples with labels $\mathcal{D}_{train} := \{(x_i, y_i) \mid x_i \in \mathbb{R}^d \wedge y_i \in \{-1, 1\}\}$, where the label y_i indicates the presence of a target concept c. In the machine learning and pattern recognition literature [DHS00, Bis07] different models have been introduced and suggested. However, in the following the most frequent and prominent ones in the context of concept detection are briefly outlined:

Support Vector Machines (SVMs) A popular choice in supervised learning are Support Vector Machines [SS01, Vap00], which are used in most concept detection systems nowadays [SW09]. SVMs are founded on linear maximum-margin classification i.e. they search for an optimal hyperplane, which serves as a decision boundary between two classes (represented by their labels $y_i \in \{-1, 1\}$). This optimal hyperplane separates samples $x_i \in \mathbb{R}^d$ from the corresponding classes such that their distance from the hyperplane is maximized. This distance is called the *margin*. The second fundamental element of SVMs is their use of *kernels*. As in many practical classification settings, a linear separation of the given samples is not achievable, SVMs map each sample x_i into a potentially high-dimensional space $\mathcal{H}$ using a projection function $\Phi : \mathcal{R}^d \to \mathcal{H}$. An additional advantage of SVMs is their ability to abstract from the space $\mathcal{H}$ by the pairwise calculation of the kernel (or similarity) $K(x_i, x')$ for all training samples. This property is known as the *kernel trick*. Typical kernel functions are the linear kernel, RBF kernel and the χ^2 kernel [SS01].

More formally, let the hyperplane be determined by the normal vector $\mathbf{w}$. Then the hyperplane can be defined as a linear combination of training samples on the margin (called *support vectors* x_i):

$$\mathbf{w} = \sum_i y_i \alpha_i x_i \tag{2.1}$$

where the coefficients α_i are the solution to the following quadratic optimization problem with linear constraints in its dual form:

$$\max_{\alpha_i} \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.2}$$

for any $i = 1 \ldots n$ and is subject to the following constraints:

$$0 \leq \alpha_i \leq C \quad \wedge \quad \sum_i \alpha_i y_i = 0 \tag{2.3}$$

where C determines the cost of misclassifying training samples and is usually assumed to be optimized as a free parameter. Ideally, in a probabilistic setting, the classification result would be a class posterior probability. However, in the case of SVMs the output of classification is the (signed) distance of a sample x_i to the hyperplane $\mathbf{w}$, which is seen as a score. To provide posterior probabilities these scores can be transformed by a sigmoid fitting into a posterior probability [Pla99].

Maximum Entropy Another example of discriminative classification is Maximum Entropy [NLM99]. Similar to SVMs it models a decision boundary between classes. However, while a SVM follows a margin maximization approach, Maximum Entropy methods learn a boundary such that the class posterior is as uninformative as possible. The method has been applied to concept detection on video [ABC$^+$03] and images [JM04, LSW10, LGS$^+$11]

Nearest Neighbor (NN) One further, quite straightforward approach to classification is nearest neighbor matching [DHS00]. This method labels an unknown sample by the labels of the most similar training samples. Similarity is defined as the distance in feature space and matching is often done by finding K of such nearest neighbors. This approach performs well in combination with large amounts of labeled training data [TFF08].

Neural Networks In contrast to the above approaches, neural networks [RHW86], as inspired by the mammal brain, model classification as a network of multiple neurons organized in a layered architecture. These neurons serve as firing functions, and learning in these networks is understood as parameter optimization of edge weights between neurons. A prominent training algorithm is the back-propagation of errors between the output class labels and the input features of the network. Neural networks have been applied in visual learning in the context of multi-task learning [Gon08] or parsing natural scenes [SLNM11]. Recent developments on the structure of such networks appear to be promising. In particular deep learning of recursive neural networks proved to be successful in large-scale visual recognition tasks [KSH12].

Decision Trees Finally, the last of the presented classification methods are decision trees [Qui86]. Decision trees utilize a tree-like graph or model of decisions along a feature vector of a sample. Learning in such context is realized by the construction of a binary tree from class-labeled training tuples. A prominent extension to decision trees are random forests [BZM07] but also other types such as bagging and boosted trees exists [Die00]. Decision trees and their random forest extensions have been used in computer vision [SJC08, CSK11], human pose detection [RRR+08], and medical image analysis [KGZC13]

Alternative approaches for classification in the context of visual learning include discriminative online learning [PUB09], kernel discriminant analysis [C. 09], generative mixture models [CCMV07], or topic models [MGP04, FFP05].

2.5.5 Intra-concept Fusion

Commonly, concept detection systems utilize multiple types of features and supervised learners [ABC+03, GMH+08, SW09], each providing a separate output for the same target concept (see Figure 2.3). To obtain concept detection results it is required to combine each single detection output into one signal. Such a *fusion* can be performed on two different levels: *early fusion* (or feature fusion) [ABC+03, GMH+08] , i.e. the concatenation of feature vectors prior to detector learning, or *late fusion* (or classifier fusion), where detector results are combined after classification [JNY07, SWS05, WLL+07]. While the former offers the advantage of utilizing feature dependencies, the latter does not have to deal with an increased high dimensional feature vector (curse of dimensionality problem [DHS00]). Please note that the merging of keyframe-based detection scores, shot-based motion detection scores, or timeline-based audio detection scores is usually referred to as *synchronization* [SW09].

2.5.6 Concept Relation Modeling / Inter-concept fusion

A final step in a concept detection processing pipeline is the semantic relation modeling between concepts [ABC+03, QHR+07, NJW+09, SvdSF+13]. The idea is to exploit the fact that the presence of

a concept serves as an additional indicator for a related concept, e.g the presence of a "road" indicates an increased probability of a "car" being present and but reduces the probability of the concept "boat-ship". Such co-occurrences or correlation between concepts can be modeled with different approaches such as *learning spatial models* or *learning temporal models*[SW09]. A similar idea is the addition of external knowledge such as in the form of taxonomies like WordNet [Fel98], ImageNet [DDS+09], or ontologies [HSSV03, NST+06]. Orthogonal input clues have also been utilized in the refinement of concept detection scores such as the video clip audience's demographic group information [UKB12, UBK13].

2.5.7 Pipeline Configuration

As seen above, a concept detection pipeline can be very complex, with each component offering a set of different approaches and methods to choose from [SWWdR08]. Since all described methods have their inherent strengths and limitations it is suggested to configure a concept detection pipeline according to given system requirements such as concept vocabulary, data quality, data quantity, and processing demands. In this case the *configuration setup* becomes a matter of selection from the pool of pre-processing methods (shot segmentation, keyframe extraction), feature descriptors, classification models and post-processing methods (fusion, concept relation modeling). In this context the constructor of a concept detection pipeline must decide on the number of different feature-classifier combinations to be used and – to make the configuration setup more multifaceted – consider parameter optimization issues of supervised machine learning [DHS00, FB04]. In fact the learning taking place to create a statistical classifier is further challenged by the usually limited amount of training samples and the simultaneous goal to optimize classification parameters such that the best possible generalized performance can be achieved, i.e. the classifier's capability to detect unseen samples of the same concept during application phase. Reasons for poor generalization can be two-fold, either by the misalignment of feature descriptor dimensionality and the amount of training data, commonly known as the *curse of dimensionality* or an extensive optimization on the training data, known as *overfitting* [SW09].

To tackle this problem several schemes have been presented in concept detection. A pioneering and successful scheme is the one introduced by IBM Research [ABC+03, NTYS07, CCC+11, BCC+13]. At each stage of the analysis it selects the best of multiple alternatives based on system performance on validation data. Similarly, the MediaMill system follows its "best-of-selection" [SW09] scheme selecting the best performing of all available paths through the system setup, which is – again – achieved with the help of proper validation data [SW05, SWvG+06a, ea11, SvdSF+13, MGvdSS13b]. Another proposed scheme focuses on feature diversity and subsequent parameter optimization such as the systems of Tsinghua University [CLL+06, WLL+07]. Despite the used schemes in concept detection an alternative to such configuration questions can be *meta-learning* [VD02, RSD12b, RSG+12].

2.6 System Evaluation

This section provides an overview of the most common evaluation methodologies and performance metrics in concept detection. Due to the nature of video with its vast amount of data and copyright issues involved, evaluation of this medium was difficult in the early days of video analysis. Empiric evaluation – however – is necessary to make concept detection systems comparable and exchange best practices

within the research community. This situation changes after the introduction of several evaluation and benchmarking campaigns such as the MediaMill Challenge [SWvG+06a], VideoCLEF [LNJ10], or its extension benchmark MediaEval [LSE+12] (not to mention established image benchmarks such as PASCAL Visual Object Classes [EVGW+08], or ImageNet's Large Scale Visual Recognition Challenge [DDS+09]). In fact, the most significant benchmark and de facto standard in video evaluation is NIST TRECVID SIN [SOK04, SOK06, OAM+13]. It provides shared data, common evaluation metrics and a platform for sharing resources among researchers. The following introduction in evaluation methodology and performance metrics is based on TRECVID standards, as commonly accepted by the research community.

2.6.1 Methodology

Concept detection system performance is driven by experimental evaluation. As previously mentioned, this is needed for the comparison of system setups but also for the finetuning of system configuration (Section. 2.5.7). An essential element in every experimental evaluation is a *ground truth* i.e. the annotation of video clips with concept labels. A ground truth is usually acquired manually and is therefore rare and demands a huge effort [AQ08]. Once, a sufficient ground truth is available for a dataset $\mathcal{D}$, common procedure is to split the dataset into three disjoint sets, $\mathcal{D}_{train}$ a *training* set, $\mathcal{D}_{test}$ a *test* set and $\mathcal{D}_{valid}$ *validation* set with $\mathcal{D} := \mathcal{D}_{train} \cap \mathcal{D}_{test} \cap \mathcal{D}_{valid} = \emptyset$. The general setup is therefore to train classifiers on the training set and to optimize system parameter in conjunction with the validation set without consideration of the test set. This set of sample is saved for the final evaluation of the entire concept detection system. With this setup overfitting is minimized and the capability for generalization of the detector is maximized. It is highly recommended to never mix training and test sets or to optimize system parameter on the test data. This is considered bad practice in the research community and is sometimes also negatively called "training on the test data". Since ground truth is rare, it is however common practice to employ techniques such as *cross-validation* during parameter optimization. In favor to increase the amount of representative samples for the given target concept, the training and validation set is split into n folds and in each optimization iteration $n-1$ fold are taken for optimization (i.e. training with different parameters) and the remaining fold is taken for the validation of performance.

2.6.2 Performance Measures

Once ground truth is available a concept detection result can be taken and a performance measure can be calculated accordingly. Because of the intersection between information retrieval, computer vision, and machine learning different performance metrics were established for system evaluation. Commonly used performance metrics in concept detection have therefore been adapted from either a *retrieval task*, where the goal is to rank a list of relevant documents according to a query or a *classification task*, where the goal is to assign a concept label to an individual document.

Following the information retrieval point of view, let $L_c \subseteq \mathcal{D}_{test} \wedge L_c^k = \{x_1, x_2, \ldots x_n\}$ be the detection result of length n and with ranks k for the target concept c, this is e.g. a ranked list of keyframes, shots, or videos in the test set. Further, let $R \subseteq \mathcal{D}_{test}$ be the set of all relevant samples in the test set, then *precision*, the number of relevant documents in the result, is defined as $p = |L \cap R|/|L|$ and *recall*, the number of relevant documents from all the available relevant documents in the test set is defined as $r = |L \cap R|/|R|$. A perfect retrieval result would be $L = R$ with precision and recall at 1.0. Another

metric is the combination of both previous measures is called the harmonic mean or *F-Score* $= \frac{2*p*r}{p+r}$. These are three common metrics from the information retrieval area. Since precision and recall influence each other it is also common to plot precision-recall curves illustrating their relation in the context of detection results and relevant samples in the test set. However, one of the most widely used metrics to evaluate relative video retrieval systems – and also the one used in TRECVID – is *average precision* (AvgP) [VH+05] or its derivative *inferred average precision* (Inf-AvgP) [YA06]. Average precision is a single-valued measure that is proportional to the area under a recall-precision curve. This value is the average of the precision over all relevance judgments in L. At any given rank k let $R \cap L^k$ be the number of relevant samples in the top k of L. Then AvgP can be defined as:

$$AvgP = \frac{1}{min(|R|, n)} \sum_{k=1}^{n} \frac{R \cap L^k}{k} \psi(x_k)$$

where indicator function $\psi(x_k) = 1$ if $x_k \in R$ and 0 otherwise. It can be seen that AvgP favors highly ranked relevant results. When a concept detection system consists of more concepts to evaluate $|Voc| > 1$, the mean of all AvgP for each individual concept is taken to indicate the whole system performance. This measure is called *Mean Average Precision* (MAP).

For classification evaluation, different metrics are utilized. Usually a result contains samples and their assigned concept labels. According to the true label of the sample, a differentiation between *true positive* (tp), *false positive* (fp), *true negative* (tn), and *false negatives* (fn) can be made. From these values different error rates such as *False Positive Rate* (FPR) or *True Positive Rate* (TPR) can be derived. Similarly as before these two measurements can be put into relation with a plot called *Operator Receiver Curve* (ROC). The single measure of such a curve plot is the *Area under Curve* (AUC). Another view on error rates for classification-based evaluation is the *Equal Error Rate* (EER), which can be used to find the optimal threshold for a system balancing equally between False Positive Rate (FPR) and False Negative Rate (FNR).

2.7 Label Acquisition Characteristics

As already outlined, one particular problem of concept detection is its demand for labeled training sets, which serve as a foundation for supervised machine learning, the underlying technology of current concept detection systems. So far, training samples were acquired manually, i.e. a human operator labels videos or video shots with respect to concept presence. Thereby, concepts are well defined according to a concept vocabulary [NST+06]. This distinct difficulty is visible in TRECVID's 2011 collaborative annotation effort [OAM+11], a joint attempt by the research community to acquire labels for the TRECVID campaign. In 2011 the TRECVID benchmark aimed to increase the vocabulary size from 130 to 500 semantic concepts for its Semantic Indexing Task (SIN) and video material had to be annotated with the new concepts by voluntarily participating groups. During a 6 week period and an involvement of 34 groups worldwide, each providing $30k - 45$k annotations, the collaborative annotation effort collected around 4.2 million labels from the given 400h of video material (consisting of 266k individual shots to be inspected). Estimating a time demand of 2sec. per annotation, the complete annotation workload adds up to $2,333$h, which can be translated into 1.3 work years[2]. Unfortunately, even with this vast amount

[2]commonly known estimation parameter of 8 hours being 1 day and 220 days being 1 year

of work put into the annotation effort, the goal to increase the concept vocabulary size of 500 concepts could not be achieved. The final concept vocabulary for TRECVID 2011 SIN task was comprised of 346 concepts, many with a minimum of 4 positive annotations per concept. Please note that the annotation effort was employing active learning methods to boost the discovery of positive annotations [AQ08].

This rather disappointing result illustrated that this time-consuming and cost-intensive effort – although leading to high quality training material – suffers from a scalability problem [SWvG$^+$06a, YH08a, USKB10] and points to the demand for alternative sources and label acquisition schemes for concept detector training. In recent years the use of socially tagged web images and video as alternative sources of training data for semantic concept detection has gained traction [USKB08a, UKSB08, BKUB09, SS09, UKBB09]. Utilizing such data gives the following advantages over training from a small set of expert labels. First, it allows to learn large concept vocabularies which are required to cover users' information needs and thus lead to a more efficient search [HYL07]. Second, it enables concept detection systems to be more flexible in learning new emerging concepts like "Sochi Olympics 2014", "Royal Wedding" or "Edward Snowden". Third, it prevents overfitting as learning from only a small set of sample videos tends to deliver detectors that generalize poorly [YH08b].

Web video is publicly available on a large scale from online portals like YouTube, Vimeo or Blinkx and is associated with a noisy but rich corpus of tags, comments and ratings that are provided by large communities. Utilizing this information might replace expert labeled datasets by automatically harvesting training material from the web. For example, to learn a concept like "person playing soccer" a search query has to be formulated and sent to one of the previously mentioned web video portals. The resulting list of relevant videos can now be downloaded and used as training material. For this purpose tags are used as positive labels for concept learning. Web video has already been proven to train more general detectors performing better on unseen datasets as compared to detectors trained on specific expert labeled data [USKB10] and demonstrated its potential as a comprehensive training source for visual concept learning [UKBB09].

On the other hand, does it suggest to entirely focus on user-generated tags and neglect effort done in the direction of visual learning of semantic concepts? Unlikely, as the following study which was done in the context of this thesis reveals. During a time span of 6 months (September 2011 – March 2012) *blank videos* were crawled every day from YouTube. The crawling process was explicitly searching for all videos only uploaded on that particular day and with the following queries: '*.mpg", "*.mov","*.avi", "img*". After the retrieval of the result list, each video was double-checked tht it had been not assigned with any meaningful title, description, or tags. A typical video retrieved by this setup can be seen in Figure 2.7 (left). In total about 108k videos were retrieved and their meta-data was stored for later investigation. After a further period of 3 months (June 2012), the meta-data of each video in the entire list of videos was re-checked and compared to the initial upload status. The intention of the study was to estimate how many people on YouTube tagged their videos after the initial upload. Please note that during the 6 month retrieval phase only videos with no tags, no description and no meaningful title were kept in the list. A comparison between the initially uploaded meta-data and the re-checked meta-data could be done for 86k videos out of the 108k. The reason for not being able to access the remaining videos on YouTube were meta-data errors ($\sim$ 8k), removed video ($\sim$ 10k), and changed privacy permissions ($\sim$ 4k). Interestingly, only 18% of video owners changed titles, 9% changed a video's description and only 11% added tags after uploading videos to YouTube. A distribution of how many videos have how

Figure 2.7: **Left:** Although popular videos are tagged well on YouTube, there exist a long tail with many sparsely tagged videos or videos with no tags, description nor meaningful title (red box) at all. **Right:** An evaluation about the tagging behavior of YouTube users for freshly uploaded videos. The graph plots the number of videos with no tags (left end) to the number of videos with up to 50 tags (right end). As can be seen, the majority of videos do no have any tags assigned or are sparsely annotated.

many tags (with 0 tags starting from the left and up to 50 tags at the right end of the spectrum) is shown in Figure 2.7 (right). As seen, the majority of videos is tagged sparsely or not tagged at all.

Summarizing, this study demonstrated that there is a high demand for the visual learning of semantic concepts, which was also highlighted by Google itself [ATY09]. Nevertheless, YouTube can serve as a reliable source for well-tagged training videos for concept detection (a simple search for the initial example concept "person playing soccer" returns over 645,000 videos on YouTube).

Web Video Characteristics Web video, when used as a training source for concept detection has its own characteristics. In particular, the usage of tags as concept levels in the context of machine learning has to be rethought. A major focus of this thesis is to adapt the different levels of labels as introduced in Chapter 1 in the context of distinct visual recognition tasks. First, web video is known to be of a very dynamic nature with over 100 hours of new video content being uploaded to YouTube every minute [YOU13]. Are these video clips reflecting real word events or are they focusing entirely on non-relevant user-generated content? To be more specific, is web video on YouTube correlated with current trending topics and correspondingly can such tagged video clips be used as labels for trending topic specific detector training (Chapter 3)?.

Second, user tagged web video has – when compared to expert labeled material – a differently motivated labeling. Experts annotate videos according to well defined concept definitions and independent of their personal interest, whereas web users strongly follow the focus of interest [UKBB09] i.e. such labels may be of a subjective nature showing non-relevant content. This behavior is often also referred to as *framing of user intent* [KL09, HKL12]. Additionally, the retrieval of training data is usually done through a search engine query consisting of a set of keywords delivering a list of relevant videos and secondly the download of those videos. How to formulate a query to receive the labeled video clips for

training and how are tags aligned within video stream (Chapter 4)?

Finally, can user-generated tags be utilized to learn more complex concept structures such as Adjective Noun Pairs? Such combinations of adjectives and nouns do not only identify concept presence of nouns but also encode the subjective understanding of adjectives. This subjectivity may differ from person to person. For example, one person may upload an image of a dog and tag it: "dangerous dog", whereas another person may not perceive the depicted dog as dangerous. To this end, it is crucial to show if the majority of such adjective noun pairs convey the perception of a large enough group to represent "dangerous dogs" comprehensively enough. Moreover, since this mid-level representation of visual content is utilized to learn more abstract labels, such as positive or negative sentiment it is necessary to demonstrate its capability to grasp sentiment in general (Chapter 5).

Chapter 3

Dynamic Vocabularies by Trending Topics Discovery

Today's concept detection systems are centered around the notion of fixed concept vocabularies not being aligned with users' information demands. To overcome this problem, this chapter presents the idea of dynamic vocabularies to synchronize concept detection with ongoing real world events. This is accomplished by mining social media for *trending topics*, which are either mapped to a fixed concept vocabulary or trained as individual concept detectors on demand. The key contributions of this chapter are[1]:

1. A system is presented that mines three major social media channels for trending topics. During an observation period of an entire year, this system performed autonomously a clustering and re-ranking of 40,000 potentially overlapping candidates to identify 2986 individual trending topic providing insights about their life-cycle and multi-channel behavior.
2. A novel fully automated trending topic forecasting approach is introduced. The approach is based on a nearest neighbor forecasting technique exploiting the assumption that semantically similar topics exhibit similar behavior.
3. In experiments on a large-scale dataset of Wikipedia page view statistics this forecasting is shown to be superior to other methods achieving a mean average percentage error starting with 45% for one day forecasts to 19% for 14 days forecasts (n=22,400).
4. A novel approach for concept vocabulary expansion is presented, which allows to dynamically add trending topics to the concept vocabulary by either linking them to a static concept vocabulary, by a direct visual learning of trends, or by augmenting the vocabulary with trending topics.
5. It is demonstrated in experiments on 6,800 YouTube clips and the top 23 target trends that by a visual learning of trending topics, improvements of over 100% in concept detection accuracy can be achieved over static vocabularies (n=78,000).

These results allow us to conclude that concept detection can be extended to dynamic vocabularies providing systems which are synchronized to current real-world events.

[1]This chapter is based on the authors' work in [BHK+09, BKUB09, UKBB09, BUB11b, BUB12, BL12, ABHD13]

3.1 Introduction

With the the growing proliferation of images and videos over the last years, the demand for multimedia retrieval tools has increased. Here, *concept detection* [SW09] – the automatic recognition of objects, locations or actions – offers the possibility of a content-based semantic search, which is of particular interest for web-based services like YouTube hosting huge amounts of weakly labeled content [YOU13]. An open issue with concept detection is the selection of suitable *vocabularies* of target concepts. These are usually picked manually by experts [NST+06], according to a given application domain (e.g. news broadcasts) or academic purpose (e.g. for performance benchmarking). One problem with this approach is that concept vocabularies are difficult to scale and adapt, which limits the applicability and suitability of concept detection (or systems built on top) to deal with the enormous diversity of web-based video. Instead, concept vocabularies should evolve when new topics of interest arise in the user community.

Simultaneously, as the consumption of multimedia content is rapidly increasing, social media streams or channels capture with remarkable accuracy what people currently pay attention to and how they feel about certain topics. Such topics are usually associated with a subject (i.e. a textual label) and often experience a sudden spike in popularity ("trending"), which often relates to real world events such as sports highlights (Olympics 2012), product releases (iPhone), celebrity news (Steve Jobs' death), disasters (Sinking of the Costa Concordia), political movements (Occupy), or entertainment (Academy Awards). By detecting these *trending topics* for detector training, concept detection can be tailored to the latest user interests. As proposed information source for trend detection, this work suggests to utilize the large variety of social multimedia services available to users on the Web (e.g Twitter, YouTube, Facebook, Flickr, Google, and Wikipedia). These sources reflect different user needs such as information demand, social communication, as well as sharing and consumption of multimedia content creating a heterogeneous multi-channel environment within the social media landscape.

As illustrated in Figure 5.3, the presented approach in this chapter aims to automatically detect trending topics, prioritize them by forecasting their impact, and add the most promising ones dynamically to the vocabulary of a concept detection system by using them as input for visual detector training. Minor and ephemeral trends – where detector training would be less rewarding due to the forecasted short life time – are neglected. This way, concept detection is able to provide dynamic concept vocabularies with on-the-fly trained detectors being aligned with latest user interest. During the investigation of the trend-based evolution of concept detection vocabularies using YouTube as application domain, the following three key questions are addressed:

1. **Is it possible to extract the "right" trends from social media?** The system mines Twitter posts, Google searches, and Wikipedia access statistics for trending topics and shows that the resulting trends are strongly correlated with YouTube uploads. This indicates that other social media channels do form a reliable indicator of user interest in the web video domain i.e. if a trend emerges on other media, video uploads on YouTube spike correspondingly. Furthermore, an analysis about trending topic life-cycles and cross-media relationships is performed investigating their behavior and coverage among various topic categories.

2. **As a trend emerges, can we predict its significance?** To adapt concept detection to trends, it is essential to identify high-impact topics and train visual detectors "on the fly". To do so, the challenge of *forecasting* the life cycle of a trending topic is addressed, i.e. predicting the

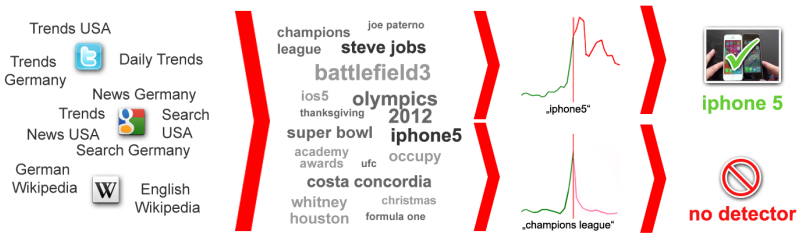


Figure 3.1: Concept vocabularies that automatically evolve with user interest: Trending topics are discovered by social media mining (left), prioritized by forecasting their impact (center), and visual detectors are trained for the most prominent trends (right).

amount of user engagement towards it at the very moment it emerges. For this purpose a novel forecasting approach is presented, which exploits the assumption that semantically similar topics exhibit similar behavior in a nearest neighbor framework. Such topics are uncovered by mining topics of similar type and category on DBPedia [ABK+07], a structured representation of the information on Wikipedia. The proposed fully automatic approach is evaluated with hundreds of trending topics on a large-scale dataset of about 317 billion views of the roughly 5 million articles on Wikipedia.

3. **Are "trend detectors" more accurate than static concept vocabularies?** The last question is about the utility of adding trending topics to concept vocabularies and whether adding these pays off in terms of detection accuracy. To answer this question, two general detection strategies are investigated, namely *(i)* linking a targeted trending topic like "Olympics 2012" with pre-trained concepts like "Athletics" or "Stadium", or *(ii)* training a new "Olympics 2012" detector as the trend emerges and videos tagged with it are uploaded. Both strategies are compared and a combination of both – *(i)* and *(ii)* – is presented that merges trends into the concept vocabulary.

Results on 6800 YouTube clips (541 hours of video) show that the 23 most prominent trends from Winter 2011/12 could be detected with much higher accuracy using trend detectors instead of a static concept vocabulary of 233 concepts. Summarizing, the contribution in this chapter is three-fold. It presents (1) a trending topic detection, which automatically mines trends from major social and online media channels, (2) an automatic forecasting technique for these trending topics based on a nearest neighbor approach exploiting semantic relationships between topics evaluated on a large-scale Wikipedia user behavior dataset, and (3) a novel approach for dynamic vocabulary expansion for visual concept detection, which allows either to map emerging concepts to an already available concept vocabulary, train directly on demand, or combine both strategies together for best detector performance.

This chapter is organized as following: First an overview of work related to trending topic detection, forecasting, and the use of vocabularies in concept detection is outlined (Section 3.2). Second, the discovery and multi-channel analysis of trending topics is presented (Section 3.3). In addition, the trending topic forecasting approach is described in Section 3.4 and a dynamic construction of concept detection vocabularies is introduced in Section 3.5. A discussion concludes this chapter (Section 3.6)

Figure 3.2: A visualization of the trending topic clustering for Wednesday, 12th Oct 2011. It can be seen that the number of cluster members is driven by the retrieved raw data. Further as depicted, the selection of cluster labels is able to find meaningful descriptions for each cluster.

3.2 Related Work

The review of related work is divided into four parts. First, work in the area of large-scale topic and event discovery is outlined. After that, research about the combination and analysis of signals from multiple media channel is reported. Then, efforts in context of forecasting behavioral dynamics are presented and finally research in the area of vocabularies in concept detection is outlined. Related work about concept learning in general is skipped since it is covered in Chapter 2 in full detail.

3.2.1 Topic and Event Detection

Detecting and tracking topics in news media has been a field of study for years [All02]. Its initial goal was the understanding of broadcasted news in multiple languages and across multiple media channels (including television and radio sources). Therefore first challenges included the segmentation of audio-visual news streams into individual story units, the identification of emerging topics in these streams, the tracking of stories that discuss a particular topic, and the determination of the original story that mentions a new topic for the first time [CSG+02]. Recently, topic and event detection has regained traction because of the availability of large datasets from social media. While prior work on trend discovery focuses on blogs and Twitter content [GHT04, KLPM10], current research focuses on Twitter, which has gained much attention recently in the area of online event detection. Due to the characteristics of Twitter and its vast amount of tweets every minute this is non-trivial rending it a very challenging task [WL11]. Weng and Lee tackled this challenge by performing a wavelet analysis on the frequency-based word signals to detect new events and further cluster terms via a graph partitioning technique [WL11]. In contrast to Weng and because it is also known that Twitter can be noisy and only a partial view on the entire database can be given as reported in [KLPM10, BNG11] approaches exists that employ aggregated trends provided by platforms like Twitter itself [KLPM10, CL09] or extract trending topics only from a subset of tweets [BNG11].

Similar to the former, this chapter utilizes lists of trending topics provided by platforms and further process and aggregate those trending topics to groups of real-world events. Additionally and different from related efforts, this chapter employs a diverse set of (textual) news streams to identify trending topics across channels and over time. This allows the presentation of a multi-channel analysis of trending topics over an observation period of one year.

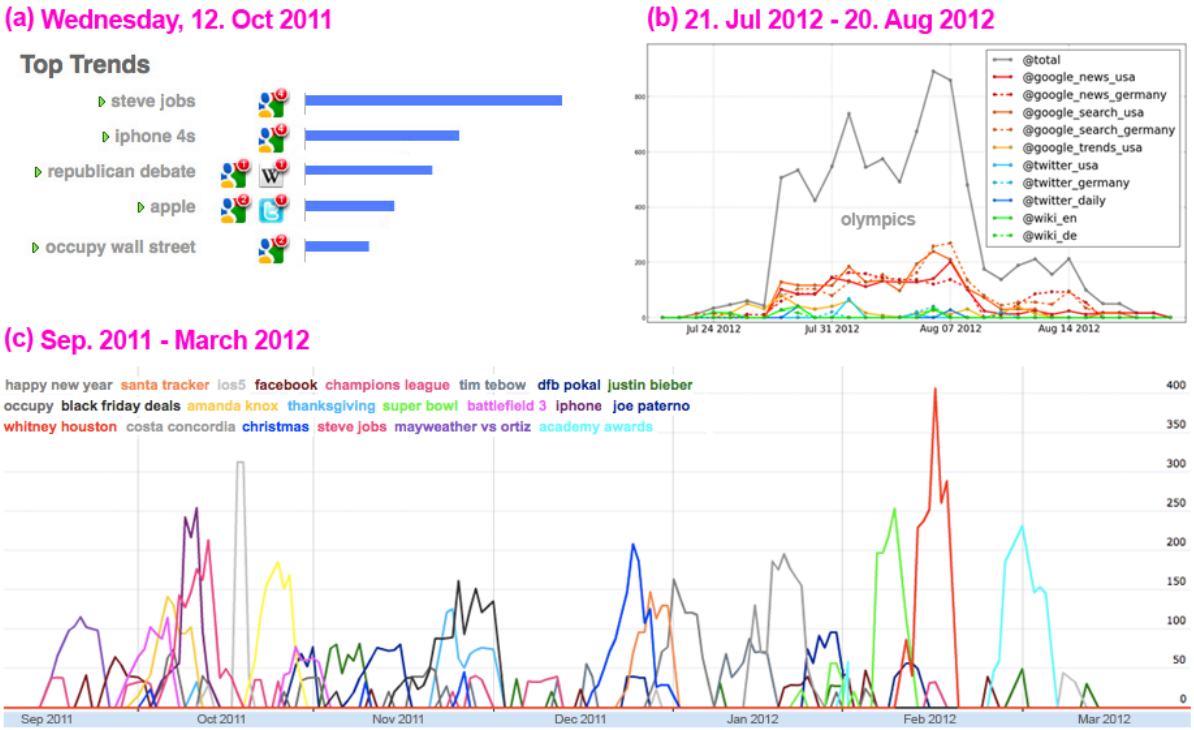


Figure 3.3: **(a)**: For each day, top trends are discovered by aggregating feeds from Google, Wikipedia, and Twitter, and trend *scores* are computed. **(b)**: The trending topic "Olympics 2012" during summer 2012. The colored curves represent the contribution of the different online and social media channels. **(c)**: The trend scores for the 23 most prominent trends, from September 2011 until March 2012.

3.2.2 Multi-Channel Analyses

Motivated by the initial idea of topic tracking between multiple channels such as news broadcast and radio, the analysis of trending topic behavior over multiple media channels is of interest for this work. Ratkiewicz et al. [RFM10] argue that although the dynamics of short-lived events such as news are relatively well understood, online popularity of specific topics (e.g. "Barack Obama") in general cannot be characterized by the behavior of individual news-driven events (e.g. "Barack Obama inaugurated as U.S. President") since the former might subsume many different news stories making it difficult to differentiate between them. They further find bursts in Wikipedia traffic to be correlated with bursts of Google search volume indicating a sudden increase of attention on the Web at large. Wikipedia article views have also been correlated with behavior on Twitter in order to analyze the potential of creating new content for breaking news (e.g. Japan earthquake) or vice versa updating pages at the moment of news (e.g. Oscar winners) [WA12]. Such correlations can also be utilized to filter spurious events on Twitter. This however indirectly resulted in the finding that Wikipedia lags behind Twitter by about two hours as measured by the textual similarity between tweets and "bursting" Wikipedia articles [OPM+12]. Adar et al. correlate several behavioral datasets such as general queries from an observation period of one month from MSN and AOL search logs with the general idea of understanding past behavior to predict

future behavior [AWBG07]. Although not focusing on trending topics, this multi channel investigation provided an analysis of the temporal correlation within topic clusters as well as differences in popularity and time delays between sources. A similar behavior was observed in [YL11], where six distinct temporal patterns with regard to certain phrases or memes have been identified to describing the rise and fall of user attention in weblogs trailing mainstream media by one or two hours for most of the considered phrases.

In contrast to the above, this work explicitly focuses on trending topics in online and social media channels over a long observation period. Similar to [DCCC11] using trending topics such as "oil spill" and "iPhone" to evaluate tweet selection, this work utilizes trending topics for semantic concept selection. It additionally investigates multi-channel behavior, analyzes the correlation between topic categories and media channels and proposes a fully automatic approach of forecasting trending topics in terms of future impact in the context of attention as opposed to visualization tools [AWBG07] or predicting cluster assignment [YL11].

3.2.3 Forecasting Behavioral Dynamics

Much research has been devoted to predicting economic variables such as auto sales or unemployment claims [CV12] by "nowcasting" them from online observations, or opening weekend box-office revenue for movies [GHL+10]. In the same way, popularity of online content has often been treated as a single variable (e.g. total number of views) instead of a time series [SH10]. Either early popularity [SH10], or content-based features such as publisher, subjectivity, or occurrence of named entities [BAH12] are used to forecast eventual popularity. A different approach is taken by Radinsky et al. who predict the top terms that will prominently appear in future news (such as prediction "oil" after observing "dollar drop") [RDM08]. More recent work treats the popularity of queries and clicked URLs of search engines as time series and uses state space models adapted from physics for forecasting [RSD+12a]. With respect to forecasting Wikipedia article popularity, the study of Thij [TVLK12] has to be mentioned. This analysis – however – restricts itself to featured articles on the main page only and accounts for daily cycles in viewing behavior.

Please note that trending topics lead to time series characteristics with unexpected shifts. These shifts are known as structural breaks in the field of econometrics and can lead to large forecasting errors and unreliable forecasting models [CH09]. A good example for such a structural break is Whitney Houston's death in February 2011 that caused 2000 times (!) more people to access her Wikipedia article than usual. Such structural breaks also described as parameter non-constancy are the main cause of large, unexpected forecast errors in practice [CH09]. Autoregressive models (AR) and extensions incorporating moving averages, seasonality and exogenous inputs [HK08] as usually utilized for forecasting, lack the necessary robustness to deal with structural breaks as introduced by trending topics. Pooling or combining forecasts from different models has often been found to outperform individual forecasts and to increase forecast robustness.

In the presence of nonlinearities, Nearest Neighbor techniques for forecasting have been found to improve out-of-sample forecasts [FRSRAF99]. Forecasting the progression of trending topics in the very moment they emerge is different in the sense that it requires a fully automatic system and that often there is little historical information available (unlike for many economic variables of interest). For example, few people were aware of "Costa Concordia" before the ship sunk in January, 2012. Similarly, to make

predictions about the "54th Grammy Awards" one needs to understand the relationship of this event with previous ones such as previous instances of the Grammy Awards. This work assumes that semantically similar topics share characteristics and behavior and therefore could improve forecasting accuracy. This assumption has not yet been explored in previous work (e.g. [RSD+12a]). In addition, the proposed approach can forecast popularity of arbitrary topics (represented by Wikipedia articles) that exhibit very diverse viewing dynamics.

3.2.4 Social Multimedia Applications

The analysis and forecasting of trends also has intersections within the multimedia domain. One application employing social media in conjunction with multimedia systems has been explored in [JGC+10]. This work uses the Flickr photo upload volume of specific topics to inform autoregressive nowcasting models for monthly political election results and product sales (i.e. the model requires the Flickr upload volume at time t to produce a forecast for time t). Also, the Flickr queries relevant to the forecast subject of interest are chosen manually (e.g. using "Hillary" instead of "Clinton" to avoid images by Bill Clinton for the 2008 Democratic Party presidential primaries). SocialTransfer is another system that uses trending topics obtained from a stream of Twitter posts for social trend aware video recommendation [RMZL12]. Learning new associations between videos based on current trends is found to be important for improving the performance of multimedia applications, such as video recommendation in terms of topical relevance and popularity. The popularity of videos is also used in [WSC+12] to drive the allocation of replication capacity to serve social video contents. This work analyzes real-world video propagation patterns and finds that newly generated and shared videos are the ones that tend to attract the most attention (called temporal locality). They further formulate the challenge to estimate the videos' popularity for video service allocation for which they use the number of microblog posts that share or re-share the video. These insights are incorporated into the design of a propagation-based social-aware replication framework. Two other research prototypes that seek to enhance the multimedia consumption experience by extracting trending topics and events from user behavior on the Web are SocialSensor [DPK+12] and TrendMiner [SPPC+12]. The first one emphasizes the real-time aspects of multimedia indexing and search over multiple social networks for the purpose of social recommendations and retrieval. The other – TrendMiner – focuses on real-time methods for cross-lingual mining and summarization of large-scale stream media and use cases in financial decision support and political analysis.

3.2.5 Vocabularies for Concept Detection

Typically, vocabularies of concept-based video retrieval systems [SW09] contain a wide range of semantic concepts such as objects, location and activities to be detected in video streams. These vocabularies are expert-defined, where visual discriminability, utility for retrieval and availability of training data have been identified as important characteristics of "suitable" concepts [NST+06]. The Large-Scale Concept Ontology for Multimedia (LSCOM) [KHN+06] is such an concept vocabulary balanced according to these criteria. It consists of $1,000$ concepts carefully selected with respect to their usefulness for news video retrieval. Although restricted to one particular domain, LSCOM served over the last year as foundation for many concept detector systems like University of Columbia DVMM's and University of HongKong VIREO's ones, consisting of a subset of 374 trained detectors from LSCOM [YCKH07,

Table 3.1: Top 30 international trending topics during September 2011 – September 2012. There is a wide variety of trending topics including sport events, product releases, celebrity news, incidents, political movements, and entertainment. Please note that the US presidential election was in November 2012 and is therefore not listed here.

	Topic		Topic		Topic
1	olympics 2012	11	christmas	21	iphone
2	champions league	12	steve jobs	22	happy new year
3	iphone 5	13	manhattan	23	kindle
4	whitney houston	14	academy awards	24	ncaa brackets
5	mega millions numbers	15	formula 1	25	em 2012
6	closer kate middleton	16	justin bieber	26	amanda knox
7	facebook	17	joe paterno died	27	earthquake
8	costa concordia	18	battlefield 3	28	mayweather vs ortiz
9	black friday deals	19	muammar gaddafi dead	29	santa tracker
10	superbowl	20	ufc	30	thanksgiving

JYCN08, NJW+09]. Another set of large vocabularies is defined by TRECVID [OAM+12]: starting with single digit vocabulary sizes in the beginning of the campaign to today's 346 concepts in its Semantic Indexing task (SIN). The origin of the SIN vocabulary is a subset of 500 concepts merged from LSCOM [NST+06] and CU/VIDEO vocabularies [JYCN08]. Roughly at the same time the MediaMill group defined their concept vocabulary of 101 concepts based on manual inspection of the TRECVID 2005 corpus [SWvG+06a]. This effort was further extended by MediaMill to a vocabulary of 500 semantic concepts [SWH+06]. In the context of video collections for concept detection, the Heterogeneous Audio Visual Internet Collection (HAVIC) [SMF+12] has to be mentioned. The collection is primarily focusing on multimedia events (up to 75 multimedia events have been defined) which can be found in thousands of hours of video material. The distinctiveness as compared to the above vocabularies is the construction of *event kits*, a complex definition of events with textual description, explication, and evidential description. One significant result of all these efforts was the experience how time-consuming annotation according to defined concept from the vocabulary of such datasets is. Nevertheless, although the costly acquisition of training data [AQ08] poses a limiting factor to vocabulary size, large-scale concept sets such as ImageNet [LLZ+11] or Google's Video2Text system [ATY09] exist.

The work presented in this chapter bears similarities to the above mentioned work in a sense that web video is used as a domain and that user-generated tags together with their video content are exploited as a source for training and the testing of concept detectors "on the fly". The key difference, however, is that this work combines web-based concept detection with trending topic discovery to develop *dynamic vocabularies* that adapt to evolving user interests. Evolving tag vocabularies have also been studied in [DJLW07], where an inductive transfer was applied upon a fixed black-box vocabulary to adapt to a users' personalized tagging behavior over time. The presented approach differs from previous ones as it trains new concept detectors based on a discovery of trending topics over large user communities, which – to the best of my knowledge – has not been investigated before.

Table 3.2: Trending Topics Dataset Overview. Statistics of analyzed trending topics and sequences for Google (G), Twitter (T), and Wikipedia (W) and combinations thereof.

	# Topics	# Sequences
Total	200	516
Google (G)	191	445
Twitter (T)	118	232
Wikipedia (W)	69	108
G & T	115	174
G & W	66	86
T & W	43	57
G & T & W	42	48

3.3 Trending Topics Detection & Analysis

This section introduces the dataset used for the analysis of trending topics across media channels and presents an analysis of their temporal characteristics as well as insights into the relationship between channels and topic categories.

3.3.1 Trending Topic Discovery

The discovery of *trending topics* i.e terms that experience a spike in user popularity is done on three major online media channels namely by analyzing posts on Twitter, statistics of Google searches, and Wikipedia site accesses. These are clustered to account for different spellings and paraphrases, and finally aggregated across time *and* channels to obtain *trend scores* describing their overall popularity. The entire process is performed automatically by the Lookapp for Ads system [BL12].

Raw Trending Topic Sources

As outlined above, Google, Twitter and Wikipedia are used as a starting point by retrieving a ranked list of popular terms from 10 different sources on a daily basis: five Google channels (Search and News for USA and Germany as well as the Trends feed), three Twitter channels (daily trends for USA and Germany as well as the Daily Trends stream), and two Wikipedia channels (popular articles in the English or German language). For each of these feeds a ranked list of 10-20 topics (in total 110 topics per day) is retrieved. Such lists of raw trending topic strings might contain multiple variations of the same entity ("occupy wall street" and "occupy") or different spellings ("Yulia Tymoshenko" and "Julia Tymoschenko"). Also, different styles of naming entities per channel exist. While Wikipedia is a very *clean* channel immediately providing a URI identifying the named entity, Twitter, in contrast, is a very uncontrolled channel riven by its hashtag system of tagging tweets. In total this type of raw topic crawling results in 40,000 potentially overlapping topics for the observation period of September 2011 - September 2012, which are covered in this dataset.

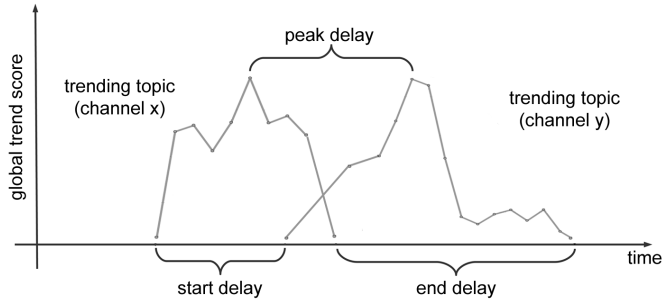
Unification and Clustering of Trends

To make use of the raw data multiple instances of the same trending topics must be connected across time and media channels. This is accomplished by a unification of terms i.e. the linkage of terms to named entities and a subsequent clustering of the individual entities on top of this. First, a mapping of individual topic strings to a corresponding Wikipedia URI is performed by selecting the top-most Wikipedia result of Google search for that topic (this approach was found to be more robust than more direct methods on Wikipedia). As a result, for each topic in the lists a topic string / URI pair is made available for the next step, the clustering of individual items from the initially retrieved lists. During the clustering two items are clustered together bottom up, if their their Levenshtein distance of their topic string or Wikipeida URI is below a certain threshold (set to 0.35 $\times$ word length) . The method allow to unify topics such as "super bowl time", "super bowl 2012", and "superbowl" into a single cluster. Cluster label assignment is done according to the topic string of the highest ranked cluster member item. Figure 3.2 illustrates such a clustering for Wednesday, 12th October 2011. It can be seen that the number of cluster members is driven by the input of raw trending topic lists. Further, as visualized in the figure, the selection of cluster labels is able to find meaningful descriptions for each cluster as compared to other potential cluster members. Overall this procedure leads to 2986 clusters or individual trends. This is only a fraction of 13% from the overall set of candidates, which have been identified as distinct trending topics over channel and time. A cluster is now represented by its most prominent member and will be referred to as a *trending topic* for the rest of this work.

Ranking Trending Topics

To reason about the popularity of trending topics for each trend a score is assigned. This assignment is based on the following method: For each day and for each of the 10 feeds, the rank at which a topic appears in its list is recorded. These ranks are combined using Borda count, obtaining a score for each day that is assigned to the topic's cluster (Figure 3.3 (a)). This Borda count ranking given clusters a high rank if its cluster members are also ranked at the top of their retrieved trend lists. Obviously, one single trending topic may emerge in multiple media channels (Figure 3.3 (b)). This fact will be of use during the lifecycle analysis across multiple channels. As seen for the trending topic "Olympics 2012", all media channels are involved with different contributions to the overall trending topic progression. To measure the impact of a trending topic over its overall lifetime, its global trend score has to be defined. This is realized by taking the sum of its daily scores over the observation period (Figure 3.3 (c) for the first half of the measured timespan). The top trends with respect to this global trend score are shown in Table 3.1. This ranking serves as a foundation for the selection of trending topics as discussed in the next section.

Some trending topics such as "Champions League" appear multiple times within the one year observation period. To allow for a life-cycle analysis of particular topics, trending topics are divided into multiple sequences being non-zero for at least two out of three adjacent dates, i.e. compensation for "score gaps" of at most one day is employed. Using this process the top 200 trends (based on their global trend score) are split into 516 (trending topic) sequences. These sequences will be used to evaluate the forecasting procedure described below.



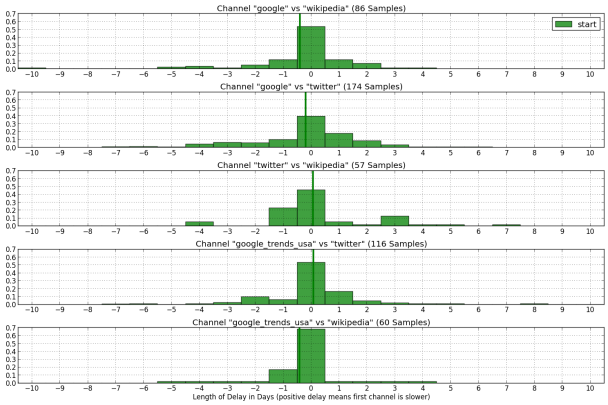
	Twitter	**Wikipedia**
	start / peak / end	start / peak / end
Google	-0.20 / 2.23 / 3.64	-0.41 / 2.35 / 4.16
Twitter	–	0.07 / 0.36 / 0.21
Google Trends	0.09 / 0.04 / 0.41	-0.42 / -0.32 / 1.25

Figure 3.4: Top: An illustration of delay calculation among media channels. Bottom: Mean delay in days between pairs of media channels (start/peak/end). Positive delay means that the "row channel" is slower than the "column channel".

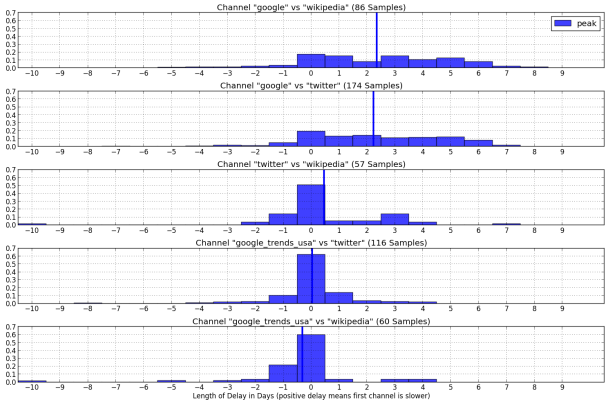
3.3.2 Lifetime Analysis of Trending Topics

Trending topics experience different amounts of attention during their lifetime. Some trending topics appear and disappear after one or two days such as "Valentines Day", some last for up to 30 days such as the "Olympics 2012", while other like "Champions League" appear multiple times within the given one year observation period. To allow for a life-cycle analysis of individual topics, they are divided into multiple sequences such that non-zero values are observable for at least two out of three adjacent dates, i.e. non-observable scores for at most one day are compensated. Please note that some trending topics might fall below the threshold of the raw trend list provided by the platform but might re-appear in that list the next day. The analysis is performed in all channels of the dataset for the top 200 trending topics (based on their global trend score). Following the previously described procedure those trending topics have been split into 516 sequences. Table 3.2 summarizes the resulting number of topics and sequences for the different channels and their combinations. Note that mapping of the ten individual monitored channels is done according to their respective sources i.e. Google, Twitter, and Wikipedia. The Google channel has the largest coverage of the trending topic sequences in the dataset (86.2 %). About 9.3 % of trending topic sequences occur in all three source channels and between 11.0 % and 33.7 % occur in two of the three source channels

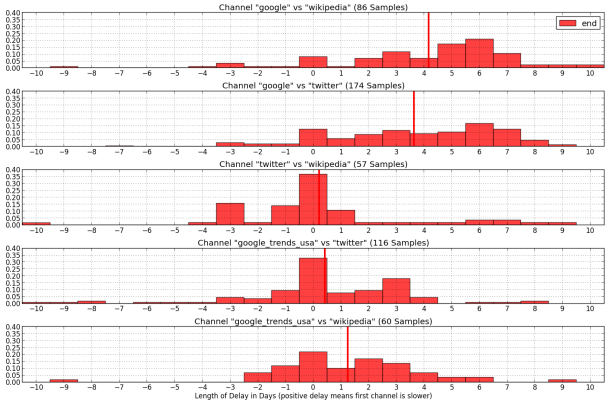
First the question of a trending topic's average time of survival is answered and whether there are differences for its lifetime in the different media channels. For this analysis, lifetime is defined as the number of consecutive days with non-zero trend scores. Histograms of the lifetime of trending topics are shown in Figure 3.6. It can be observed that the trending topics in the given dataset rarely survive longer than fourteen days (with some exceptions such as "Olympics 2012") with most trending topics having a lifetime of less than nine days. Since Google covers a large share of top trends, the distribution for the channel looks very similar to the overall distribution. The lifetime of topics on Twitter is much shorter



(a) Histogram of start delays in days. Note that Twitter is not significantly faster than other channels.

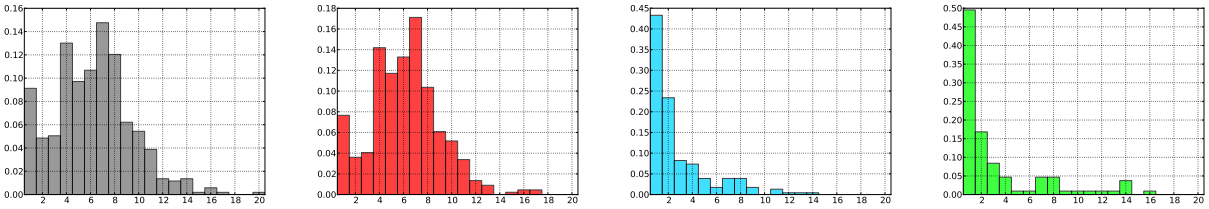


(b) Histogram of peak delays in days. Note that Google peaks after Twitter and Wikipedia in most cases. However, Google Trends USA behaves similarly to Twitter and Wikipedia.



(c) Histogram of end delays in days. Note that trends in Google survive much longer than in Twitter and Wikipedia but that this behavior is weaker within Google Trends USA.

Figure 3.5: Histogram of delays in days in the different media channels. The colored vertical bars represent the mean of the distributions.



(a) Trending Topic lifetime within all source channels.

(b) Lifetime for Google

(c) Lifetime for Twitter

(d) Lifetime for Wikipedia

Figure 3.6: Histogram of the top 200 trending topics and their lifetime of the appearance in respective social media channels in days. Due to the large contribution of Google to the overall amount of trending topic sequences its distribution looks similar to the distribution for all source channels. Interestingly, a similar distribution for the Twitter and Wikipedia source channels can be observed

in accordance with the expectations of the ephemerality of trends in this channel, about two thirds of the top trends only survive for one or two days. Interestingly, the Twitter distribution looks similar for Wikipedia providing a first indication of similar behavior of trending topics in these two channels.

For trending topics that occur in at least two of the three channels an analysis of their behavior in these channels can be performed. To this end, three points in time are of particular interest to capture their characteristics: the day on which the trend starts, the days on which the trend peaks (defined by the trend score defined earlier), and the day on which the trend ends again. Comparing two channels with each other, the *delay* between those channels can be defined as the difference between the start/peak/end dates in channel X and the start/peak/end dates in channel Y. Note that a positive delay means that the first channel X is slower, i.e. trends tend to start later in channel X than in channel Y. The mean delays for start, peak, and end are summarized in Table 3.4. Interestingly, there are only marginal differences in starting delays (first of the three numbers) between the three channels with Twitter and Wikipedia being slightly faster than Google. These results are not very surprising considering that Osborne et al. found Twitter to be around two hours faster than Wikipedia [OPM+12] – a difference almost impossible to observe in data of daily granularity. A much stronger effect is observed for peak and end delays (second and third number). Both Twitter and Wikipedia tend to peak more than two days before Google. The picture is even clearer when looking at the end delays where Twitter and Wikipedia lead Google by three and four days respectively. Overall, these results add to the ephemerality of the Twitter and Wikipedia yielding the second indication of similar characteristics of these two source channels.

However, this analysis indicates that some of the Google channels seem to be intentionally delayed or averaged, i.e. showing the top stories over an average of several days. As a control for this Google Trends can be used since it is known to not be artificially delayed. As seen in the delay analysis it peaks around the same time as Twitter and Wikipedia, and its trending topics tend to end around half a day after Twitter and more than one day after Wikipedia.

3.3.3 Cross-Media Topic Category Analysis

To provide insights about what kind of topics are the most popular in the individual channels all 200 top trends have been manually annotated with categories. The categories were chosen by examining the

Table 3.3: List of categories and their associated trending topics. Note that a trending topic might be assigned to multiple categories.

Category	Description	#Seq	Examples
sports	sports events, clubs, athletes	52	olympics 2012, champions league, bayern muenchen, superbowl, eli manning
celebrity	person with prominent profile	49	steve jobs, kim kardashian, michael jackson, neil armstrong, justin bieber, whitney houston
entertainment	entertainers, movies, TV shows	39	grammys, emmys, heidi klum
politics	politicans, parties, political events, movements	32	paul ryan, occupy, christian wulff, paul ryan, gauck, kim jong il, occupy, acta, muammar gaddafi dead
incident	an individual occurrence or event	27	costa concordia, hurricane isaac, virginia tech shooting, aurora shooting, reno air crash
death	death of a celebrity	22	whitney houston, joe paterno died, neil armstrong
technology	product or event related to technology	20	iphone 5, ces, nasa curiosity, ipad, space shuttle, google+, battlefield 3, apple, higgs boson
actor	actor in TV show or movie	18	lindsay lohan, michael clarke duncan, bill cosby
product	product or product release	15	ipad, windows 8, diablo 3, iphone 5, kindle
artist	music artist	15	justin bieber, miley cyrus, beyonce baby
holidays	day(s) of special significance	11	halloween, thanksgiving, valentines day
company	commercial business	10	apple, facebook, megaupload, instagram
show	TV show	7	x factor, wetten dass, the voice
movie	a motion picture	6	dark knight rises, hunger games, the avengers

main themes found in the dataset. Please refer to Table 3.3 for more information about the individual categories, their descriptions, and examples. Since a trending topics such as the death of "Whitney Houston" might match multiple categories such as "celebrity", "entertainment", "death", and "artist", an individual trending topic might be assigned to multiple categories. The engagement with respect to the different categories in a media channel is measured as follows: For each trending topic within the channel its score is assigned to all of its categories. Finally, the scores for each channel are normalized, e.g. to account for the dominance of Google for the scores overall. The resulting distribution over categories is displayed in Figure 3.7.

It can observe that channels have a tendency to specialize in certain topic categories. For example, the most popular category in Google is sports. A large share of the scores is also assigned to celebrity and entertainment categories. Google also has the highest relative share (15 %) for politics. Twitter also features many trends in the celebrity and entertainment categories. Interestingly, it has the highest relative shares of trends related to products, companies and technology. One reason might be that a large fraction of Twitter are technology affine early adopters that like to share their thoughts on new products. Another interesting finding is that over 20 % of the scores on Twitter are assigned to the holidays category. A hypothesis is that holiday related trends are big on Twitter as many people tag their posts and pictures with the same hashtags such as #christmas or #thanksgiving. Wikipedia clearly shows a specialization for categories that involve people and incidents such as disasters or the death of celebrities. Contrary to the intuition that Wikipedia is a slowly evolving channel which people use to read up on complicated topics, especially when also considering the temporal properties of the Wikipedia channel from the analysis above, it can be said that many users use Wikipedia for additional information

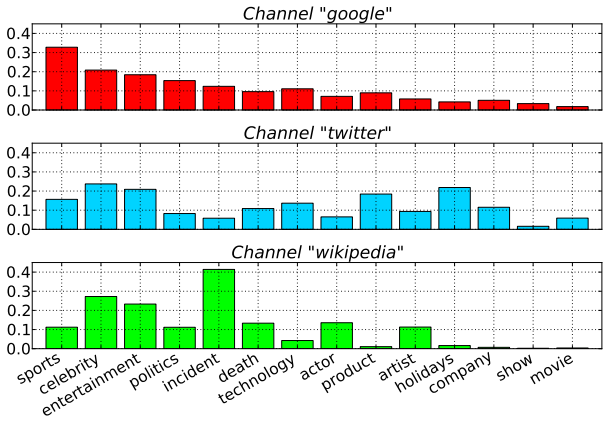


Figure 3.7: Normalized distribution of trending topic scores over trend categories in the individual channels. Note that channels tend to specialize in certain categories, e.g. Twitter for product related topics and Wikipedia for incident ones.

about these trends and events to learn about or remind themselves about related topics.

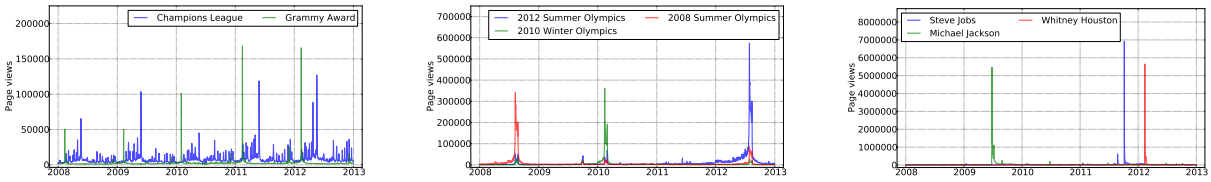
3.3.4 Classes of Pattern Recurrence

During the analysis of different trending topics, three major classes of signals were identified. Examples for all three classes are given in Figure 3.8.

- **Class 1 - Self-recurrent** The first class of signals exhibits recurring patterns within the same signal. These *self-recurrent* signals can be seen in Figure 3.8 (a). The self-recurring behavior of the trending topic "Champions League" is given due to the yearly scheduled soccer competition with breaks during the winter and summer and the finals in May.

- **Class 2 - Recurrent** The second class of signals does not exhibit recurring patterns themselves but recurrence can be found within other related signals and are therefore referred to as *recurrent* signals. For example, Figure 3.8 (b) shows the trending topic "2012 Summer Olympics", being a Summer Olympics (blue). Since this signal does not have much of a history before 2012 its signal pattern does not feature any attention before 2012. However, previous Olympics such as the "2010 Winter Olympics" and the 2008 Summer Olympics, reveal similar behavior of having two peaks, one at the start and one at the end of the event indicating the opening and closing ceremony. Note that this is not a simple yearly seasonality as the Summer Olympics happen every four years or is time shifted as in case of Winter Olympics. In such cases, where instances of the same real-word event are available recurrence is given. Unfortunately, the rules to find such instance naming rules (naming, numbering) may become arbitrarily complex as for e.g. the "Super Bowl" using the Roman counting system (e.g., Super Bowl XLVI).

- **Class 3 - Non-recurrent** The third class captures *non-recurrent* signals which do not exhibit recurring behavior and for which there is no obvious related or preceding instance. An example is given in Figure 3.8 (c) which shows different trending topics about celebrity deaths. Obviously, such events only occur once so any form of self-recurrence is impossible.



(a) Case 1: Self-recurrent signal (b) Case 2: Recurrent signal (c) Case 3: Non-recurrent signal

Figure 3.8: There are three different classes of behavioral signals with large implications for forecasting: (a) Self-recurrent, (b) recurrent, and (c) non-recurrent signals.

After having a first understanding of trending topic life-cycles, multi-channel behavior and signal pattern, the challenging task of forecasting trending topics can be addressed as presented in the next section.

3.4 Forecasting of Trending Topics

As motivated in Section 3.1, anticipating high-impact trending topics is useful for the deployment of evolving vocabularies satisfying users' information needs. However, forecasting trending topics is a very challenging problem since the corresponding time series usually exhibit highly irregular behavior (structural breaks) when the topic becomes "trending". Pooling or combining forecasts from different models has been found to increase forecast robustness in econometrics literature [CH09].

In this section, a forecasting approach is proposed that combines time series from multiple semantically similar topics. In particular because of the seen class differences of signal recurrence and their characteristics this idea is of importance since it has a large impact on forecasting. For example, the first class of *self-recurrent* signals could be forecast with straight-forward Autoregression methods as mentioned in Section 3.2. The second (*recurrent*) and third (*non-recurrent*) class of signals is much more challenging to cope with since semantic similar topics have to be found to build a model for forecasting future progression of the signal. However, finding a particular pattern in a corpus of the size of e.g. the *entire* Wikipedia corpus (which serves as the time-series dataset for the evaluation) is clearly non-trivial as there are about nine billion patterns to choose from (assuming five years of daily page views on five million articles). Therefore, the goal is to identify semantically similar topics exhibiting similar behavior to inform forecasting.

A conceptual overview of the proposed approach is presented in Figure 3.9. On the very left the trending topic for the"Olympics 2012" (Summer Olympics) can be seen along with its Wikipedia page view statistics during 2012. The task to be solved is to forecast the number of page views for a period of 14 days (yellow area) from the day indicated by the red line. This point in time for forecasting is triggered by the emergence of a corresponding trending topic in the observed channels (see Section 3.3). Note that the time series exhibits complex behavior such as the smaller peak at the end of the forecasting period which most likely corresponds to the closing ceremony event on that day. Based on the assumption that semantically similar events can exhibit very similar behavior, the first step is to automatically discover related topics such as previous Summer and Winter Olympics or FIFA/UEFA soccer championships (as illustrated by the second box). The second step in Figure 3.9 shows patterns of user engagement for

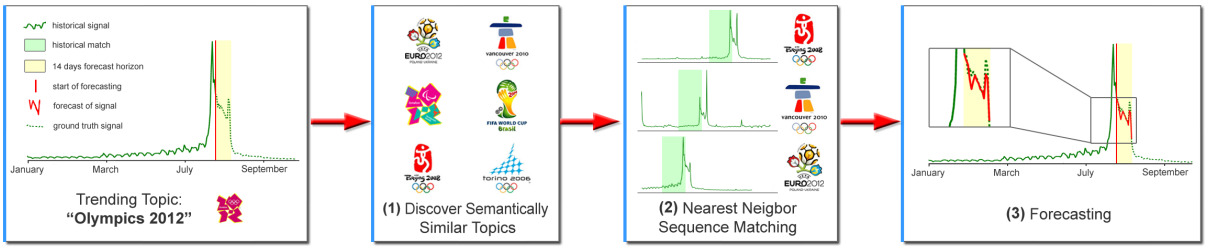


Figure 3.9: System overview of the proposed forecasting approach. For a given trending topic first semantically similar topics are discovered, then those are searched for patterns of similar behavior which are then taken to produce a forecast.

Summer Olympics 2008, the Winter Olympics 2010, and the UEFA Euro championship 2012 that were found to match the current behavior of the 2012 Olympics best (green history window). These patterns show certain similarities such as a second peak for closing ceremonies or final matches. The identified sequences from the previous step are then combined to a forecast shown at the very right. In the next section, these individual steps are explained in more detail.

Please note that this work is neither able to, nor can attempt to predict incidents such as natural disasters or sudden deaths of celebrities in advance. However, even for unpredictable events like these, the patterns of user attention once this event has happened can be forecast e.g. by taking previous instances of natural disasters or celebrity deaths into account.

3.4.1 Discovering Semantically Similar Topics

For a given trending topic, semantically related topics are discovered with the help of DBPedia [ABK+07], a database containing structured information about several million named entities extracted from the Wikipedia project. Since during the detection of trending topics each individual topic has been mapped to a named entity represented by a Wikipeida URIs (Section 3.3), it is reasonable to use DBPedia as a data repository providing rich semantic annotation such as category and type information. Given category (via `dcterms:subject`) and type (via `rdf:type`) information from DBPedia, a set of semantically similar topics was compiled that share most categories or types with a given trending topic. Essentially topics and properties form a large bipartite graph with other topics being linked to these properties. Following a set of connected topics can be found by traversing the graph and by ranking of the number of shared properties. For example, as seen in Figure 3.10, the Wikipedia URI for "2012 Summer Olympics"[2] is assigned to the categories `Sports_festivals_in_London`, `Scheduled_sports_events`, `2012_Summer_Olympics`, and `2012_in_London`. Its types include `Event`, `SportsEvent`, `Olympics`, and `OlympicGames`. For examples, the 2012 Summer Paralympics share most of the properties with the Olympics 2012.

Formally, at this stage the aim is to compile a topic set $\mathcal{T}_{sim}$ which includes all the discovered similar topics. For later comparisons, an additional topic set $\mathcal{T}_{self}$ is also assembled which only includes the trending topic itself (to simulate *self-recurrent signals*) and $\mathcal{T}_{gen}$ which contains a wide variety of general trending topics (in this case the top 200 trending topics to simulate forecasting on non-similar but popular trending topics). In the following, these sets will be referred to by the placeholder $\mathcal{T}$ and an individual

[2] `http://en.wikipedia.org/wiki/2012_Summer_Olympics`

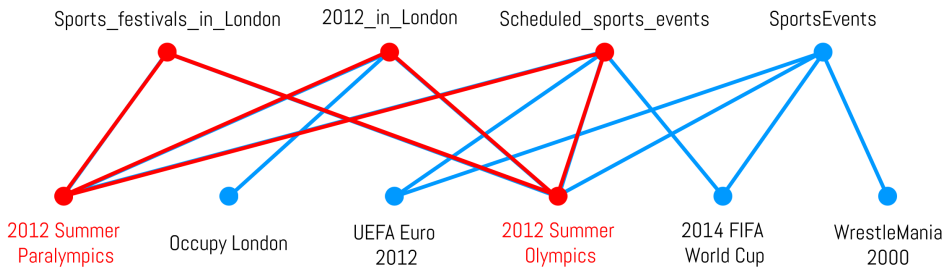


Figure 3.10: At DBPedia, the entity '2012 Summer Olympics" (red right) is linked certain semantic properties, which are also shared with other semantically similar topics (red left).

similar topic will be referred to as $\text{sim}_j \in \mathcal{T}$. An overview of the formal notation is given in Table 3.4 for reference.

3.4.2 Nearest Neighbor Sequence Matching

Obviously not all time series corresponding to the discovered similar topics look the same. Therefore, it is necessary to search within these time series for sequences that match historical behavior of the current trending topic. For example, historical topics such as the "1896 Summer Olympics" have gained very limited attention over the last years and are therefore unlikely to be representative for the large amount of engagement the "2012 Summer Olympics" experienced. To pick the right instances $\text{sim}_j \in \mathcal{T}$ from the set of similar topics, a short history window of the trending topic to be forecast is compared to each of the time series of candidates, i.e. the viewing statistics for the last two months, to all partial sequences of the same length of similar topics in the topic set $\mathcal{T}$.

To capture this step in formal terms, let $S_{\text{topic}}[t]$ be the time series for the given topic at time t. Further let define

$$S^{t_0}_{\text{topic}}[t] := S_{\text{topic}}[t_0 + t] \tag{3.1}$$

as the shifted version of the time series (used for aligning multiple series below). In the following, the forecast of $S_{\text{topic}}[t]$ is assumed to be done with a horizon of h days starting at time t_0. Given a topic set $\mathcal{T}$ from the previous step, the sequence candidate set is defined as $C(t_0)$ that includes all possible shifted time series $S^t_{\text{sim}_j}$:

$$C(t_0) = \{S^t_{\text{sim}_j} \,|\, \forall \text{sim}_j \in \mathcal{T},\ \forall t :\ t \leq t_0 - h\} \tag{3.2}$$

The condition $t \leq t_0 - h$ ensures that information more recent than $t_0 - h$ is never used in the forecasting of h days, i.e. no use of future information is allowed. Given this candidate set, the next step is to search for the k members $S^{t_i}_{\text{sim}_i}$ $(i = 1, \ldots, k)$ that are the best matches for the time series of interest $(S^{t_0}_{\text{topic}})$. Note that these nearest neighbors are already correctly aligned through shifting the time series S_{sim_i} by a corresponding t_i. Also, note that the same similar topic sim_i can occur multiple times (e.g. for repetitive signals). Formally, the nearest neighbor set becomes

$$N^k(S^{t_0}_{\text{topic}}) = \{S^{t_1}_{\text{sim}_1}, \ldots, S^{t_k}_{\text{sim}_k}\} \tag{3.3}$$

Table 3.4: Summary of the formal notation used in this chapter to describe the presented forecasting approach.

Notation	Explanation
$\mathcal{T}_{sim}$	Topic set containing all discovered similar topics
$\mathcal{T}_{self}$	Topic set only including the topic itselg
$\mathcal{T}_{gen}$	Topic set of general topics not based on semantic similarity
$\mathcal{T}$	Generic placeholder for a topic set
$\text{sim}_j \in \mathcal{T}$	Similar topic
t_0	Time of forecast
$S_{\text{topic}}[t]$	Time series for topic at time t
$S^{t_0}_{\text{topic}}[t]$	Shifted version of the time series that starts at t_0
$C(t_0)$	Sequence candidate set including all shifted time series
$S^t_{\text{sim}_j} \in C(t_0)$	Candidate sequence
$S^{t_0}_{\text{topic}}$	Time series of interest at point of time of forecast t_0
$N^k(S^{t_0}_{\text{topic}})$	Nearest neighbor set for time series of interest
$S^{t_i}_{\text{sim}_i} \in N^k(S^{t_0}_{\text{topic}})$	Nearest neighbor sequence (time series)
$d(\cdot,\cdot)$	Distance metric for time series
$\mathcal{F}(S^{t_0}_{\text{topic}})[\tau]$	Forecast for time series of interest τ days after t_0
$\alpha(\cdot,\cdot)$	Scaling function ensuring a smooth forecast continuation

where $S^{t_i}_{\text{sim}_i}$ are the k distinct elements that are smallest wrt. $d(S^{t_0}_{\text{topic}}, S^{t_i}_{\text{sim}_i})$ for all $S^{t_i}_{\text{sim}_i} \in C(t_0)$. Here, $d(\cdot,\cdot)$ is a distance metric between both time series which, in this case, only depends on a short history window of the time series. An interesting question is whether the metric should be scale invariant and in which form and to what degree. The following distance metrics are proposed:

1. `euclidean`: a squared euclidean distance: $(d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2)$
2. `musigma`: euclidean distance on normalized sequences $x'_i = (x_i - \mu)/\sigma$ (where μ, σ are mean and standard deviation estimated from the respective time series)
3. `y_invariant`: a fully scale invariant metric as proposed in [YL11] $(\min_\gamma d(x, \gamma \cdot y))$.

Section 3.4.4 compares these different distance metrics in more detail.

3.4.3 Forecasting

Even the best matching sequences identified in the previous step might not be a perfect fit for the time series to be forecast. Therefore, it is necessary to rescale the matching sequences such that each aligns with the last observed value of $S^{t_0}_{\text{topic}}$. This ensures that the forecast will be a continuous extension of past behavior. Now, the forecast $\mathcal{F}$ becomes the median over scaled versions of the sequences from the previous step $(N^k(S^{t_0}_{\text{topic}}))$. Subsequently, the range of forecasting days is represented by $\tau \in [0, \ldots, h-1]$ where again h is the forecasting horizon (usually $h = 14$). Finally, the forecast can then be formally described as

$$\mathcal{F}(S^{t_0}_{\text{topic}})[\tau] = \underset{S^{t'}_{\text{topic}'} \in N^k(S^t_{\text{topic}})}{\text{median}} (\alpha(S^{t_0}_{\text{topic}}, S^{t'}_{\text{topic}'}) \cdot S^{t'}_{\text{topic}'}[\tau]) \tag{3.4}$$

where $\alpha(S^{t_0}_{\text{topic}}, S^{t'}_{\text{topic}'}) = S^{t_0}_{\text{topic}}[-1](S^{t'}_{\text{topic}'}[-1])^{-1}$ adjusts the scale of nearest neighbor time series based on the last observed score. In practice, α is limited to an interval (e.g. $[0.33, 3.0]$) for robustness. Also, the method is evaluated using the average instead of the median for forecasts.

Opportunities & Limitations

While trying to predict the future is a very hard and sometimes even impossible challenge, in many cases user behavior follows certain patterns that allow for a certain amount of predictive accuracy. For example, knowing that the "Summer Olympics 2008" attracted up to 350k daily viewers on Wikipedia a reasonable guess would be that the maximum number lies at least this high for the 2012 Summer Olympics. Furthermore, discovering certain patterns such as increased attention during the closing ceremony can improve forecasts (as illustrated in Figure 3.9). The proposed approach relies on the availability of histories of viewing statistics (or other forms of attention such as click data). In addition, it is assumed that semantically similar topics for a given trend are available. Using DBPedia for this purpose worked for most trending topics in the given dataset. Problems included trends such as "Italy Germany" (referring to the soccer match during the EURO 2012) for which the assigned Wikipedia article (named entity) was incorrectly assigned to "Italy". Suboptimal assignments like this come with a loss of prediction quality.

3.4.4 Evaluation

In this section, first the Wikipedia page views dataset is described and then quantitative results are presented for the forecasting approach proposed in the previous section.

Dataset Description

The evaluation of forecasting is done on a large-scale dataset of page views on Wikipedia, an online collaborative encyclopedia that has become a mainstream information resource worldwide and is frequently used in academia [RFM10, OPM+12]. Reasons for this particular choice of social media channel were (a) the public availability of historical views data necessary to build forecasting models (hourly view statistics for the last five years), (b) the size of the dataset allowing a comprehensive analysis of the proposed method across a wide range of topics (over 5 million articles), and (c) previous results show that user behavior on Wikipedia (bursts in popularity of Wikipedia pages) is well correlated with real-world events [RFM10]. Although results are presented on Wikipedia the approach can be applied to any online and social media channel for which historic data is available.

The raw Wikipedia viewing statistics[3] are published by the Wikimedia foundation and obtained as hourly view statistics starting from January 1, 2008 (2.8 TB compressed in total). These logs are aggregated to daily viewing statistics (URIs that have been viewed less then 25 times on that day are dropped). It can be assumed that this does not introduce any bias since trending topics tend to accumulate view counts several orders of magnitude higher. For each day this results in approx. 2.5

[3] `http://dumps.wikimedia.org/other/pagecounts-raw/`

Table 3.5: Selected trending topics along with their nearest neighbor topics using category and type information on DBPedia (step 1). The ones chosen by nearest neighbor sequence matching (step 2) are in bold print. In some cases the topic itself can be used for forecasting, e.g. if the time series contains repetitive patterns.

Trending Topic	Nearest Neighbor Topics
2012 Summer Olympics	**2008 Summer Olympics, UEFA Euro 2012, 2010 Winter Olympics** :: 2016 Summer Olympics, 2014 FIFA World Cup, 2006 Winter Olympics
Whitney Houston	**Ciara, Shakira, Celine Dion, Brittany Murphy, Ozzy Osbourne** :: Alicia Keys, Paul McCartney, Janet Jackson
Steve Jobs	**Mark Zuckerberg, Rupert Murdoch, Steve Jobs** :: Steve Wozniak, Bill Gates, Oprah Winfrey
Super Bowl XLVI	**Super Bowl, Super Bowl XLV, Super Bowl XLIV** :: Super Bowl XLIII, 2012 Pro Bowl, UFC 119
Justin Bieber	**Selena Gomez, Kanye West, Justin Bieber** :: Katy Perry, Avril Lavigne, Justin Timberlake
84th Academy Awards	**83rd Academy Awards, 82nd Academy Awards** :: List of Academy Awards ceremonies, 81st Academy Awards
UFC 141	**UFC 126, UFC 129, UFC 124, UFC 132, UFC 127, UFC 117** :: UFC 138, UFC 139, UFC 137
Battlefield 3	**Mortal Kombat, FIFA 10, Call of Duty: Modern Warfare 2, Portal, Duke Nukem Forever** :: Call of Duty: Modern Warfare 3, Call of Duty 4: Modern Warfare, Pro Evolution Soccer 2011
Joe Paterno	**Terry Bradshaw, Joe Paterno, Jack Ruby, Paul Newman, Jerry Sandusky** :: Lane Kiffin, Donna Summer, Joe DiMaggio
Tim Tebow	**Reggie Bush, Michael Oher, Peyton Manning, Tim Tebow** :: Colt McCoy, Cam Newton
Diablo III	**Call of Duty: Modern Warfare 2, Call of Duty 4: Modern Warfare, Portal 2, Portal, StarCraft II: Wings of Liberty** :: World of Warcraft, Deus Ex, Rage
Grammy Award	**Grammy Award, Emmy Award, Nobel Peace Prize** :: Nobel Prize in Literature, BET Awards, Pulitzer Prize
54th Grammy Awards	**53rd Grammy Awards, 52nd Grammy Awards, 54th Grammy Awards** :: 51st Grammy Awards, 2012 Billboard Music Awards, 2012 MTV Europe Music Awards

million URIs attracting 870 million daily views. In total, the English and German Wikipedia features more than five million articles that can be used for forecasting. Note that while this dataset is used for the forecasting of historical time series data (actual Wikipedia viewing statistics), the multi-channel pipeline described in Section 3.3 serves as a robust trigger for trending topics detection, which initializes a forecasting.

Experiments

Experiments are structured along the three main building blocks of the proposed approach to compare design choices for the individual methods independently as presented in Figure 3.9.

Discovering Semantically Similar Topics: To begin with, the influence of discovering semantically similar topics is evaluated in two ways. First, qualitative results are presented by showing retrieved similar topics for a few trending topics. Second, an indirect evaluation is provided by comparing forecast performance (i.e. through the third step) of using semantically similar topics from DBPedia to using a general set of topics (the top 200 trending topics themselves).

A representative subset of semantically similar topics for the top trends are shown in Table 3.5. Note that the method is able to successfully identify similar events or previous instances for events like the

Table 3.6: RMSE forecasting error for the baselines, selected autoregressive models, as well as methods using only the trending topics themselves (*Self*), a general set of topics (*Gen*), or similar topics (*Sim*). The number of forecasting days of the remaining 14 day period is represented by τ, e.g. $\tau = 5$ means that five days after the topic becomes trending the method forecasts the remaining nine days.

	Method	RMSEs in 1000				
		$\tau = 0$	$\tau = 3$	$\tau = 5$	$\tau = 7$	$\tau = 9$
Baselines	naive	63.2	33.1	20.2	17.4	11.4
	linear trend	86.9	48.5	28.3	23.1	14.5
	average trend	49.3	25.9	22.0	19.9	18.3
	median trend	48.1	24.9	20.6	18.1	16.1
ARIMA	AR(1)	50.1	27.8	20.1	15.9	12.7
	AR(2)	75.1	31.7	22.6	16.0	13.4
	ARMA(1,1)	53.0	28.7	20.5	15.8	13.2
	AutoARIMA	58.9	30.7	26.9	19.5	16.7
Self	average	46.0	23.7	19.7	18.0	16.6
	average_scaled	44.6	21.9	17.6	15.5	13.8
	median	46.1	23.8	19.7	17.7	16.0
	median_scaled	44.9	22.3	18.1	15.5	14.4
Gen	average	45.7	22.9	19.2	16.1	14.1
	average_scaled	45.7	22.5	16.0	14.1	11.4
	median	41.4	21.2	17.6	15.4	12.9
	median_scaled	40.1	19.5	15.2	12.8	10.2
Sim	average	41.4	18.8	16.0	14.0	12.3
	average_scaled	39.6	17.1	13.7	11.6	10.0
	median	42.1	19.9	16.5	14.0	12.5
	median_scaled	41.0	17.9	14.2	11.5	9.8

"Olympics", the "Super Bowl", "UFC events", or the "Grammy Awards". Furthermore, the method is also able to discover similar people like similar music artists, entrepreneurs, athletes, and even people that died from the same cause (such as lung cancer for Joe Paterno, Jack Ruby, and Paul Newman). On average, a set of 95 semantically similar topics is retrieved per trending topic.

Nearest Neighbor Sequence Matching: The main design choice when matching sequences from similar topics to a short history window of the time series is the choice of the distance metric. As introduced in Section 3.4.2 three different distances are compared namely `euclidean`, `musigma`, and `y_invariant`. The experimental setup for this part as well as the forecasting is given as follows. Evaluation is done with the trending topics acquired in Section 3.3. They serve as an external trigger for forecasting, i.e. for each of the 516 sequences of the top 200 trending topics, a prediction for a horizon of 14 days is made, starting on the first day they emerged. The reasons to choose this window of 14 days is given by the maximum lifetime for most trending topics (see Fig. 3.6 (a)). To compare the distance metrics, the quality of the (first) nearest neighbor returned by this metric is measured by its similarity to the actual viewing statistics over the next 14 days (similar to the forecast setting but directly using

the nearest neighbor as the forecast). The quality of this match is captured by an error metric. In this work, the root mean squared error is used as defined by $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2}$ where A_t is the actual value and F_t is the forecast value. As a relative error metric the mean absolute percentage error (MAPE) was chosen being defined as $MAPE = \frac{100\,\%}{n}\sum_{t=1}^{n}|\frac{A_t - F_t}{A_t}|$. Note that unlike [RSD+12a] the error is measured on the actual view counts instead of normalizing by the total views for each day and presumably smoothing over large relative errors (i.e. even larger trends might only account for 10^{-5} of the daily views yielding very small error rates for virtually any forecast). This relative error metric has the advantage of being easy to interpret and to compare across different time series in contrast to absolute metrics such as RMSE. However, it can become unstable if the actual value is very small.

The results measured in MAPE indicate that although `y_invariant` can retrieve poor quality matches successfully, the invariance of `musigma` and `y_invariant` does not help in this task and the simple `euclidean` distance performs equally or better than the other two metrics. Therefore, the `euclidean` metric is employed for all the following experiments. Further, note that the best matches have between 82 and 319 % error range illustrating the high complexity of the task. More details about NN sequence matching can be found in [ABHD13].

Trending Topics Forecasting: Forecasting performance is measured by forecasting the next 14 days for each of the 516 sequences of the top 200 trending topics at the point in time when they first emerge. The approach is compared to several baselines that use a short history window of the time series itself (similar to [RSD+12a]): a *naive* forecast (tomorrow's behavior is the same as today's) and a *linear trend* based on the last 14 days. In addition the forecasts are compared to the *average trend* and *median trend* in the trending topics dataset as a baseline that includes multiple time series. Note that this average and median trend are computed from μ/σ-normalized time series since the average/median of actual view counts are actually very far from most trending topics. To still be able to compute the error for actual view count prediction the values have to be de-normalize the *average trend* and *median trend* baselines with the parameters of the time series to be predicted. Further, the performance is compared against selected autoregressive models that represent a state-of-the art technique for time series forecasting, namely AR(1), AR(2), ARMA(1,1), and AutoARIMA (please refer to [HK08] for a formal specification of these models). Note that while the following experiments are performed for different numbers of neighbors, the nearest-neighbor-based results are only reported for $k = 3$, which performed best by a small margin as compared to other values of k.

The RMSE forecasting errors are summarized in Table 3.6 which reports the results for a number of instances of the proposed forecasting approach. *Self*, *Gen*, and *Sim* refer to the different topic sets $\mathcal{T}_{self}$, $\mathcal{T}_{gen}$, and $\mathcal{T}_{sim}$ from which the nearest neighbor sequences are chosen (as described in the last section). On average, the RMSE of the best method is about 9-48k views closer to the actual viewing statistics than the different baseline methods. It can also be observed that the proposed nearest neighbor approach outperforms autoregressive models in all cases which perform roughly on the same level as the baselines. Also, notice that the fairly sophisticated AutoARIMA model performs worse than its much simpler AR(1) counterpart even though it aims to choose the best ARIMA model for the underlying data and was shown to perform well in several other forecasting competitions. This observation adds to the impression that autoregressive models (which assume stationarity; see [HK08]) are not well suited to model trending topic time series with structural breaks. Further, as illustrated it can be seen that

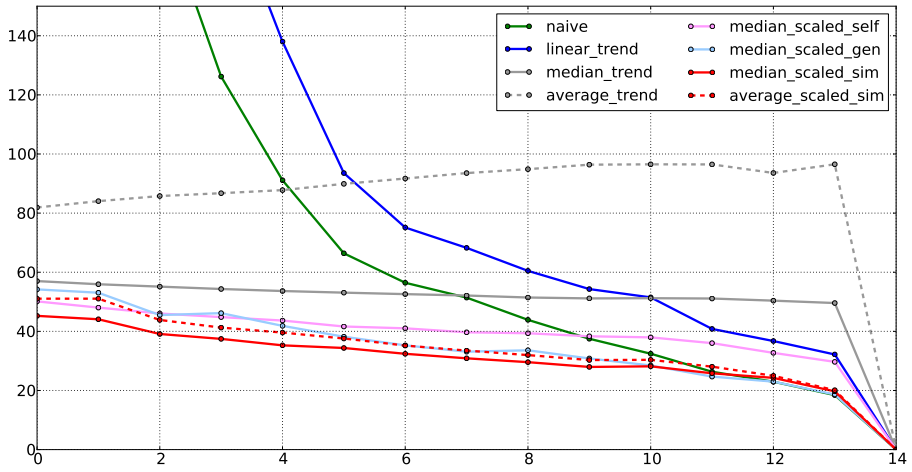


Figure 3.11: MAPE forecast error moving through a 14 day window, e.g. the error at 4 depicts the forecasting error of the following 10 days.

taking the median tends to perform about as good or better than taking the average, using scaled nearest neighbor is better than unscaled neighbors, and that using semantically similar topics (*Sim*) is better than using a general set of topics (*Gen*) which in turn is better than restricting oneself to a single time series (*Self*).

However, RMSE error has the disadvantage that it is dominated by the most popular trending topics with the largest view counts. Therefore, the MAPE measure is chosen for the remaining analysis as this relative error metric is comparable across trending topics. Additionally, because MAPE errors can become disproportionally large (e.g. forecasting 1000 views when it is actually only 100 results in a 900 % relative error), obvious outliers are dropped (5 %) and the average error is reported for the remaining sequences. The error plots only display the results for the baselines and the best performing methods from Table 3.6 to be able to visually distinguish them. The methods are evaluated by beginning with a forecast of 14 days, then a 13 day forecast after one day etc. as illustrated by the X axis in Fig. 3.11. From the plot one can observe that the proposed approach including median and scaling clearly outperforms all baselines as well as other instances of the framework. The proposed method achieves a mean average percentage error (MAPE) of 45 % for a forecast of 14 days, monotonically decreasing to 19 % for a forecast of 1 day, a relative improvement over the baselines of 90 % (14 days) to 20 % (1 day).

Example forecasts for two trending topics, “Battlefield 3” (a computer game) and “The Hunger Games” (a novel and movie), are given in Figure 3.12. Each column depicts multiple forecasts at different points in time as indicated by the vertical red line. The proposed method `median_scaled_sim` (*Sim*) is compared to its variants only using the trending topic itself (*Self*) or using a set of general topics (*Gen*). Summarizing, as already recognized by the observation of different classes of signal recurrences, utilizing only the information from the same time series (*Self*) does not provide sufficient forecasts for trending topics. In contrast, using semantically similar topics (*Sim*) leads to more accurate forecasts that e.g. are able to capture multiple peaks (such as in the Hunger Games example).

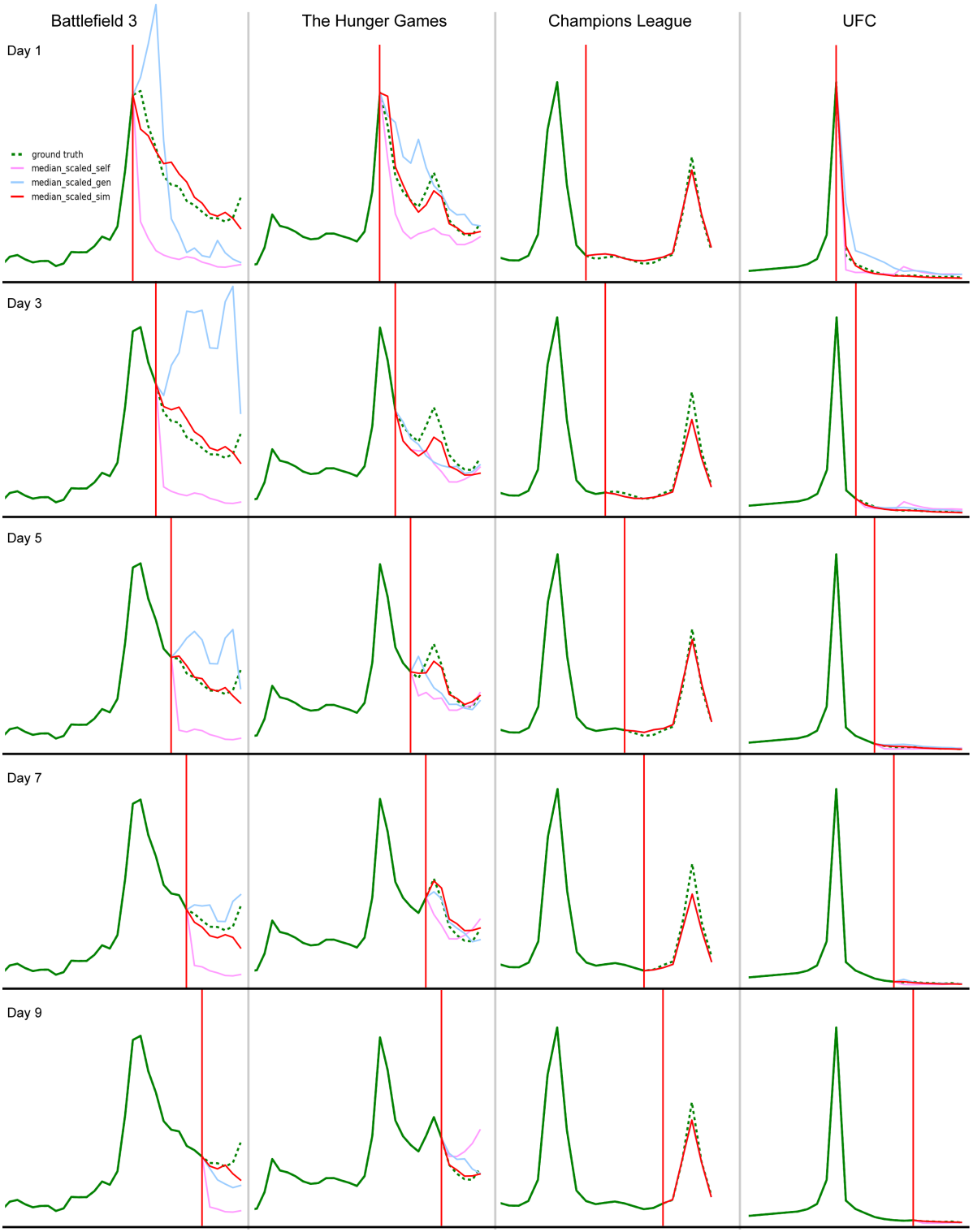


Figure 3.12: Visualization of trending topic forecasts for "Battlefield 3", "The Hunger Games", "Champions League", and "UFC". Each column depicts multiple forecasts at different points in time (as indicated by the vertical red line) for 1, 3, 5, 7, and 9 days after the trending topic emerges.

3.5 Evolving Vocabularies for Concept Detection

Based on the previously outlined approach to discover and forecast high-impact trending topics in social media channels, this section introduces the visual learning of such trends. To this end, the input stream of these trending topics serves as a set of new semantic concepts to continuously extend concept vocabularies to provide the proposed evolving nature necessary to satisfy users' information needs. In particular, this section compares visual learning of trending topics on the fly to concept detection based on static concept vocabularies.

To demonstrate the benefit of trending topics detector learning, two components are needed *(i)* a set of trending topics $t_1, .., t_m$ given by an external system and *(ii)* a concept detection system with its static vocabulary $C_1, .., C_n$ providing detection scores $P(C_1 = 1|x), .., P(C_n = 1|x)$ for a new video or keyframe (described by content-based features x). Having both components available, the goal is to estimate $P(T = t_j|x)$. Prerequisite *(i)* is given by the initially described trending topic mining and detection system *lookapp for ads* [BL12]. The second prerequisite – *(ii)* – will be outlined next.

3.5.1 Concept Detection System with a Static Vocabulary

As mentioned above one major component in the context of this chapter is the availability of a fully functional concept detection system. This system should be designed according to the state-of-the-art visual learning approach as outlined in Chapter 2 and serve a large and comprehensive set of concepts from its vocabulary Voc. Such a concept detection system is introduced in this section.

One key feature of the introduced concept detection system is its ability to train concept detectors with web video from platforms like YouTube. Detector training from web video is nowadays considered as a valid source for visual learning as it demonstrated to augment or replace traditional approaches of training data acquisition [UKSB08, BKUB09, BHK+09, USKB10]. For example, to learn the appearance of the concept C_i = "soccer" ($C_i \in Voc$), corresponding YouTube material is downloaded and used to train the corresponding visual concept detector. Once this detector is available, scores $P(C = C_i|x)$ can be computed indicating concept presence in previously unseen video content x . This way a visual learning without the tedious manual annotation of training samples is realized (please note, that Chapter 4 will go into more detail about the challenge of training from web video.

The presented concept detection system bears similarity to the *TubeTagger* [UKBB09] concept detection system with the exception that the system presented in this chapter puts an emphasis on third-party cloud computing infrastructure allowing to train multiple detectors simultaneously on-the-fly.

Concept Vocabulary The set of concepts defining the static vocabulary of the presented concept detection system are selected manually with similar intention and selection criteria in mind as done for LSCOM [NST+06]. The LSCOM was carefully constructed by manually selecting those semantic concepts which comply with the following principles: *utility* i.e. the concept should support video retrieval, *coverage* i.e. the set of concepts should cover a pre-defined domain, *feasibility* i.e. an automatic detection of concepts from the video stream should be feasible, and *observability* i.e. concepts should have been visually distinct and therefore appear frequently in the underlying video repository. These principles can be understood as commonly accepted guidelines when compiling concept detection vocabularies. Since the presented concept detection system is built upon web video as a domain, YouTube was taken

as the underlying video repository. This has the consequence, that beside known LSCOM concepts such as "airplane flying" or "boat-ship" the presented system vocabulary also contains concepts with high popularity on YouTube such as "wedding", "eifeltower", "simpsons", or "wrestling". A list of all 233 concepts belonging to the static vocabulary can be found in Appendix A.

Training Data Acquisition Prior to detector training, a set of training videos have to be acquired for each concept of the vocabulary. To retrieve such video material, a textual query is formulated and sent to the YouTube API returning a list of videos matching the query being sent. Although the returned list of matching videos is limited artificially by the YouTube API to a maximum of 1,000 videos, a single query should be sufficient to retrieve enough videos for detector training. From this list a certain number of videos is downloaded and serve as positive samples for supervised learning. Negative samples are drawn from other videos not being tagged with the concept. It is known that such query-based retrieval of videos does not always provide the right subset of suitable material for training [USKB08b, UBB10, BUB10]. In such cases to further improve the quality of downloaded material, text queries to the YouTube API were refined manually by inspecting the first YouTube result page and interactively adding additional terms and category information to the query. For example, to download training content for the concept "mountain", the query "mountain panorama" is constructed and only videos from the YouTube category "travel & places" are downloaded. An overview of refined query and category combinations can be found in Appendix A) (for more details about how to automatically retrieve suitable training material please refer to Chapter 4). In total over $50,000$ videos have been downloaded from YouTube for the entire concept vocabulary training (with at most 250 per concept). The split between training and test sets was done 50% by randomly sampling videos such that no video from the same YouTube user belongs to the training and test set.

Concept Detection System: Construction and Evaluation Considering the outlined setup as discussed in Chapter 2, the concept detection pipeline for each target concept look as follows: visual learning is performed on the basis of keyframes. Those are extracted using change detection methods directly on the video stream (in total the entire concept detection system is trained on approximately 850,000 keyframes). For each concept, a binary classification problem is formulated i.e. all keyframes sampled from videos tagged with the target concept are used as positive training samples, keyframes from other clips as negative ones. From these keyframes the well-known bag-of-visual-words representation is computed [SZ03]. This is done by sampling patches regularly at several scales (3,600 patched per frame) from each keyframe. These patches are characterized by a 128-dimensional SIFT descriptor [Low99] and matched to a 3,000-dimensional codebook of prototypical patches (the codebook was constructed by K-Means clustering). As a result for each keyframe an aggregated histogram feature is made available, which will serve as an input for classifier training. For this, a two-class SVM, which can be considered as the state-of-the-art approach in concept detection [SOK09, SW09], is trained using the LIBSVM [CL01] implementation. Parameter optimization (C and γ) is done by a grid search over a three-fold cross-validation and as a kernel function for the SVM, the χ^2 kernel is chosen:

$$K(x,y) = e^{-\frac{d_{\chi^2}(x,y)}{\gamma^2}} \tag{3.5}$$

where $d_{\chi^2}(.,.)$ is the χ^2 distance between the bag-of-visual-word histograms x and y:

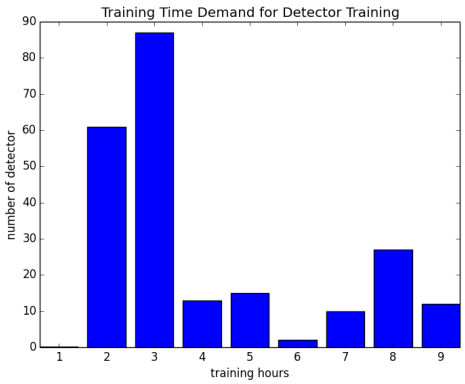


Figure 3.13: Histogram of time consumption for SVM concept detector training. It can be seen that two groups of training exist, the ones needing 2-3 hours of training and the ones needing around 8 hours of training.

$$d_{\chi^2}(x,y) = \sum_{i=1}^{m} \frac{(x^i - y^i)^2}{x^i + y^i} \tag{3.6}$$

Training data was sub-sampled to include not more that 1,000 positive samples and at most 4,000 negative ones, resulting in a slightly imbalanced dataset towards the negative class. This is aiming to represent the variability of the large concept vocabulary. Overall, the entire concept detection system with its fixed vocabulary of 233 concepts achieved a promising performance of MAP 56.7 %, which is significantly better than a detection by using random guessing (20 %).

Role of Cloud Computing Infrastructure Since video content analysis is a computational intensive task, distributed systems for high-performance computation are considered as an infrastructural key component in detector training [SGK+07]. However, employing parallel programming paradigms on huge computer clusters, which have to be built up and continuously maintained, is a costly undertaking. In contrast to such a setup, today's cloud computing services provide an alternative solution. They offer on-demand computational resources which can serve as an underlying computational platform realizing building blocks along the concept detection pipeline such as storage, parallel feature extraction and distributed classifier training.

To enable such parallel feature extraction and detector training this work uses the third-party cloud computing platform PiCloud[4] to run concept detection. Technically PiCloud provides an intermediate layer between the Amazon Web Services [5] and the Python programming language. Users of the platform can chose between multiple, so called *core type* representing different levels of computational power as being represented by the available memory and number of Amazon's "compute units" (an equivalent to a CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor). The first question in setting up such an environment for concept detection is about which core type to chose for each of the two tasks (feature extraction and classifier training). Since feature extraction is performed on keyframe level the most extreme setup would mean using one machine per keyframe. This – however – would lead to a

[4] PiCloud (`www.picloud.com`) sponsored this work with its academic research grand

[5] `http://aws.amazon.com`

waste of resources because of the overhead costs caused by setting up one machine for each of the 850k keyframes. A more reasonable setup is to split keyframes into batches of 1000 keyframes leading to a setup utilizing 850 machines. Because of its low memory consumption, feature extraction can be done on a *c2* core type (2.5 compute units, 800MB and 30GB disk space) machine smoothly. In contrast, classifier training is a more computationally expensive task demanding more memory on a single machine. Here, the *m1* core type (3.25 compute units and 8GB memory and 140GB disk space) was considered to be sufficient. Since it can be expected that each SVM training will demand its time, each SVM training representing a semantic concept from the vocabulary will be executed on a separate machine in parallel leading to a setup of 233 *m1* instances on PiCloud.

Summarizing the construction of the entire concept detection system, its feature extraction and classifier training required less than 48 hours on PiClouds cloud-computing infrastructure. When accumulated the computational time needed for feature extraction of the entire set of keyframes can be summed up to 23 hours of time costing in total the equivalent of US$ $7.0 - 10.0$ including data transfer and storage. The computational time needed for classifier training varies from detector to detector depending on the complexity of the learning task. An overview of time consumption for SVM training can be seen in Figure 3.13. The figure displays a histogram of hours for model training of individual concepts. It can be seen, that there are two groups of model training in context of time consumption. There are models which require $2 - 3$ hours of training and there are models which need around 8 hours of training. The quickest training – "badlands" and "autumn" – took 2.1 hours, the longest – "piano" – 9.71 hours on average demand 4.57 hours of computation time. Please note that this observation does correlate with the complexity of SVM training in terms of number of support vectors kept for the concept models. Altogether classifier training took 1063.22 hours or 44.30 full days of accumulated time, costing the equivalent of about US$ 320.0.

In the following a concept detection system is made available and its large vocabulary of 233 concepts can be used for a concept-to-trend mapping allowing concept detection systems to recognize trending topics visually

3.5.2 Concept-to-Trend Mapping

The baseline method to be presented is the one dealing with a static vocabulary of an available concept detection system. The idea is to take the given set of the detected concept scores and map them to estimate the presence of target trending topics t_j. For example it is reasonable to assume that pre-trained concepts from a static vocabulary like "Athletics" or "Stadium" would be able to provide meaningful clues for the trending topics "Olympics 2012". To allow for such a mapping, concept-trend similarities are estimated using the normalized Flickr distance as proposed in [JNC09]. This measure calculates a distance between two semantic concepts considering tag behavior on Flickr as a reliable indicator for concept correlations i.e. two concepts are considered *closer* if they co-occur as tags in the same images. Normalization of this measure is done by the individual counts of concept tags. Formally,

$$D(C_i, t_j) := \frac{max\{log\ count(C_i), log\ count(t_j)\} - log\ count(C_i, t_j)}{log\ M - min\{log\ count(C_i), log\ count(t_j)\}} \tag{3.7}$$

where M is the total number of Flickr images and $count(.)$ is the number of images with a particular

tag. In this case, $D(C_i, t_j)$ provides a distance for a concept $C_i \in Voc$ and a trending topic t_j, which further has to be transformed to a similarity by calculating:

$$sim(C_i, t_j) := e^{-D(C_i,t_j)/\gamma} \tag{3.8}$$

Finally, these similarities are normalized to probabilities $P(C_i = 1|T = t_j)$ (more information on the estimation of γ will follow later).

The concept detection results $P(C_i = 1|x)$ and concept-trend-similarities $P(C_i = 1|T = t_j)$ are now combined by marginalizing over all possible concept appearances as proposed in [UKB12, UBK13]:

$$\begin{aligned}
&P(T = t_j|x) \\
&= \sum_{c_1,c_2,..,c_n \in \{0,1\}} P(T = t_j, C_1 = c_1, .., C_n = c_n|x) \\
&\approx \sum_{c_1,c_2,..,c_n \in \{0,1\}} \Big[P(C_1 = c_1, .., C_n = c_n|x) \cdot \\
&\qquad\qquad P(T = t_j|C_1 = c_1, .., C_n = c_n) \Big].
\end{aligned}$$

Assuming independence of the individual concepts and applying Bayes' rule, the above formula can be rewritten as:

$$\begin{aligned}
&\approx \sum_{c_1,c_2,..,c_n \in \{0,1\}} \Bigg[\prod_{i=1}^{n} P(C_i = c_i|x) \cdot \\
&\qquad\qquad \frac{P(T = t_j) \prod_{i=1}^{n} P(C_i = c_i|T = t_j)}{\prod_{i=1}^{n} P(C_i = c_i)} \Bigg] \\
&= P(T = t_j) \cdot \prod_{i=1}^{n} \Bigg[\frac{P(C_i = 0|x) \cdot P(C_i = 0|T = t_j)}{P(C_i = 0)} \\
&\qquad\qquad + \frac{P(C_i = 1|x) \cdot P(C_i = 1|T = t_j)}{P(C_i = 1)} \Bigg],
\end{aligned} \tag{3.9}$$

whereas the priors $P(C)$ and $P(T)$ are set to uniform distributions. This way, trending topics can be estimated via concept detection and fixed vocabularies.

3.5.3 Training of Visual Trend Detectors

It can be expected that videos for particular trending topics bear similarities with certain concepts from the vocabulary. Yet, the majority of video material can be assumed to be quite specific. For example while the accident of the "Costa Concordia" matches the concepts "boat-ship" or "ocean" its visual appearance is very specific and unique for this incident.

Therefore the second strategy proposed in this chapter is the construction of trend-specific detectors from the web as described in Section 3.5.1. As a trending topics emerges, videos tagged with it are uploaded (Section 3.5.5 will discuss this condition). This circumstance is exploited and such videos are used as positive training samples (by using the presented retrieval approach in Chapter 4) to train a "trending topic detector" on the fly. Although this training set might seem as to be smaller compared to a "regular" concept detection training set, it is more focused on the target trending topics. Once

the resulting detector is available, the detector can be used to detect trending topics in other videos, estimating $P(T|x)$.

To unfold the full potential of such trend detectors two conditions are crucial. First of all, it is important to focus on trending topics, which have a particular impact i.e. the potential to accumulate attention from the general public in the near future. Obviously it does not make sense to train detectors for trending topics which are declining or are already at the end of their lifetime. Therefore it would be helpful to pay less attention to ephemeral trends since they usually disappear in $1-3$ days. Second, and closely related to this, the training of a visual detector should not take longer than the lifetime of a trending topic. Since the presented approach is already factoring out the problem of a time-consuming label acquisition the challenge, is to provide training of potential multiple trend detectors on-the-fly. Hence, the *lookapp* [BUB11b] system is proposed to deal with this issue by utilizing an on-demand cloud computing infrastructure which is able to scale virtually without limitation. Such a setup allows to train multiple trend detectors in parallel instead of delaying detector training by processing them one by one.

3.5.4 Expanding the Concept Vocabulary

Finally, a combination of the former two strategies is tested. The idea is to *expand* the available static concept vocabulary dynamically by adding the previously trained concept detector to it as a new concept C_{n+1}. Such an expansion can be of value to prevent a too strong emphasis on the trend detector. Since the training data set might be too specific additional clues in form of detection scores from a large concept vocabulary can improve the performance of trending topics recognition itself.

To make the marginalization method sensitive to the trend detector, its normalized Flickr distance is set to $D(C_i, t_j) := 0$ emphasizing that the newly added concept represents the trending topic itself and has a stronger influence on the result than other concept detection scores. The *Concept-to-trend Mapping Baseline* is then applied to the dynamically extended vocabulary.

3.5.5 Experimental Evaluation

The presented concept detection experiments cover the first half of the given observation period where 20k topics have been analyzed using the procedure outlined in Section 3.3. All trending topics have been ranked from this timespan according to their *trend score* and the top 23 ones have been picked for evaluation (see Figure 3.3 (c) for their distribution over time). Some of them are obviously very challenging to detect like “happy new year”, whereas others seem to be feasible like “battlefield 3”, a video game that was released at that time.

Correlation of Trends and YouTube Uploads

For each trend, 150 YouTube videos were downloaded (i.e. videos being tagged with the trend name). Attention was paid to carefully filter video clips outside the given 6 month test period for trending topics data acquisition. This procedure yielded a dataset with 2,500 clips (31-147 per trend). The initial idea of trend detector construction assumes the availability of video material being tagged with the emerging trending topics. Therefore first the hypothesis has to be confirmed that uploads on YouTube correlate with emerging trends. YouTube video upload for the top 10 trending topics can be seen in Figure 3.14. It can be seen in these upload histograms that upload peaks appear in conjunction with external events

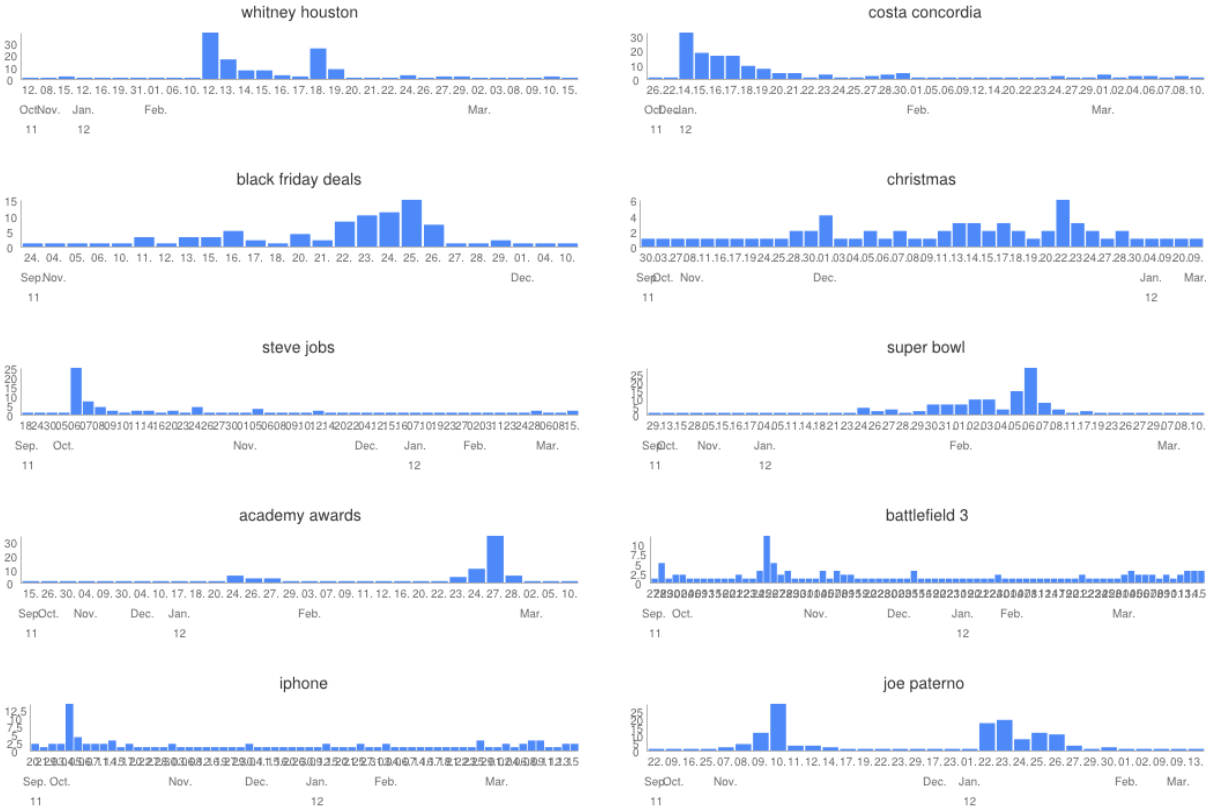


Figure 3.14: Correlation between the top 10 trending topics (first half year of observation period, Winter 2011/2012) and video upload on YouTube. It can be seen that video uploads follow external events with clear indications for peaks.

indicating a correlation with them. Speaking in terms of a quantitative evaluation, averaged over all trending topics, 57.3 % of their videos were uploaded on a "trend" day or the day after (a uniform distribution over time would correspond to 8.8 %). Thereby, event-based trends like "Whitney Houston" (referring to the death of the famous singer) display the strongest alignment between YouTube and the presented trending topics detection, while long-lasting/periodic trends like "facebook" or "champions league" the lowest. Overall, this result indicates that YouTube video uploads are closely aligned with trending topics allowing the retrieval of enough training material for "trend detector" training.

Visual Trend Detection

Next, the ability to detect trending topics in visual content will be evaluated. To do so, YouTube is queried for additional background material distributed randomly over the observation period. This material is acquired by downloading the daily *most recent* video clips with no tags. From the resulting 4,300 "background clips" and from the 2,500 "trend clips", 78,000 keyframes were extracted using the previously described change detection method. To learn the direct visual trend detectors (`Strategy 1`), a 60 %-40 % split of all clips into a training and test set was conducted. Results are reported in terms of mean average precision (MAP) on the test set (2,720 videos). As a static concept vocabulary

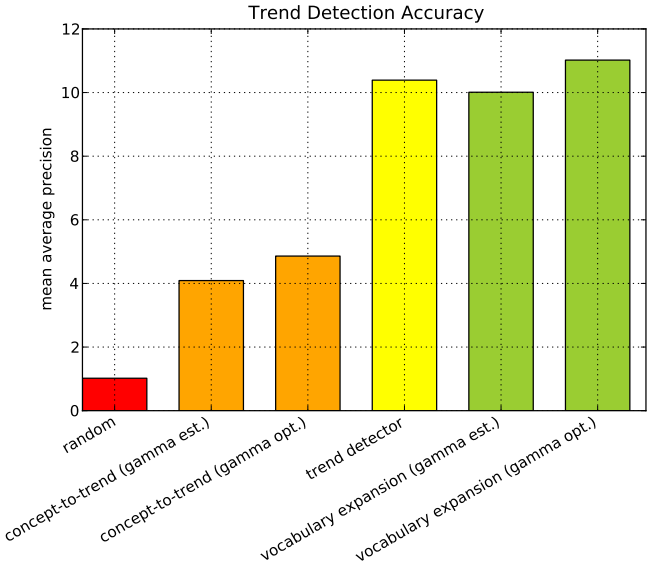


Figure 3.15: Quantitative results of trending topic recognition. A specialized trend detector (yellow) outperforms a static concept vocabulary (orange). Expanding the vocabulary with the new detector gives further improvements (green).

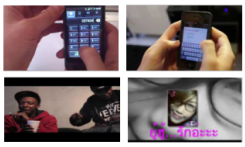
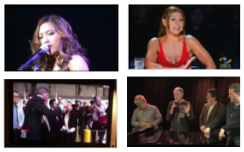
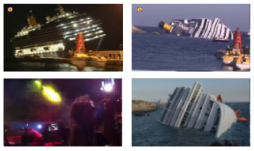
(`Baseline`), the concept system as described in Section 3.5.1 is employed, covering 233 concepts that range from "concert" over "demonstration" to "phone". Please note that these detectors were pre-trained on a held-out dataset of YouTube clips from before the observation period.

Both – concept detection and visual trend detection – are conducted on keyframe level, using visual words features (obtained by a regular multi-scale sampling of about 3,600 SIFT features [Low04], vector-quantized to 3,000 clusters using K-Means) in combination with Support Vector Machines (SVMs) using a χ^2 kernel and fitted by a grid-search cross-validation.

Quantitative results of the experiment are illustrated in Figure 3.15. A direct training of "trend detectors" (TD: yellow) performs with a MAP of 10.39 % better than the concept-to-trend-mapping (CTM: orange) with a of MAP 4.86 % and 4.09 % giving the lowest accuracy. Moreover, the vocabulary expansion strategy (VE: green bars) performs best, with an mean average precision (MAP) of 11.2 % (γs optimized by grid search) and MAP 10.01 % (estimated γ based the average of pairwise Flickr distances). This indicates that training new detectors seems a promising approach for adapting concept detection to new emerging trends, while a static concept vocabulary can help to improve accuracy further when being combined with a trend detector. If not combined with a trend detector, the stand-alone static vocabulary approach does not provide sufficient detection performance.

A closer inspection of system performance is given in Figure 3.7 for the following trends "ios5" (referring to the release of Apple's operating system), "Mayweather-vs-Ortiz" (a boxing fight), "Whitney Houston" (the death of the singer), "Battlefield 3" (a video game release), "Costa Concordia" (cruise line disaster), "Occupy" (the political movement), and "Champions League" (European soccer tournament). For each trend, the top-ranked videos for (TD) and (CTM) are displayed. Furthermore the corresponding concepts and their similarities can be seen: Some can be considered outliers (e.g. "cathedral" for

Table 3.7: The 4 top-ranked videos by the direct trend detector (TD) and concept-to-query mapping (CTM) for 3 sample trends. The last column lists the best detectors by their accuracy, including trend detectors (TD), the concept-to-trend-mapping (CTM), vocabulary expansion (VE) (γ optimized by grid search), and the best individual concept detectors.

trend	**top results (*trend detector*)**	**top results (concept-to-trend mapping)**	**most similar concepts**	**best performing rankers (AvgP)**
ios5			(1) safari (2) phone (3) cathedral	(1) TD: 43.3% (2) VE: 41.3% (3) phone: 37.1% (4) iphone: 27.3% (5) win-desktop: 25.8%
Mayweather vs. Ortiz			(1) press-conf. (2) boxing (3) rugby	(1) VE: 26.5% (2) boxing: 23.8% (3) TD: 21.7% (4) interview: 7.8% (5) wrestling: 6.9%
Whitney Houston			(1) bill clinton (2) singing (3) videoblog	(1) VE: 11.4% (2) TD: 6.9% (3) CTM: 6.2% (4) interview: 5.2% (5) obama: 5.2%
Battlefield 3			(1) tank (2) helicopter (3) soldiers	(1) TD: 28.8% (2) VE: 16.3% (3) counterstrike: 4.3% (4) fencing: 4.3% (5) car racing: 3.1%
Costa Concordia			(1) shipwreck (2) boat-ship (3) shoppingmall	(1) VE: 7.4% (2) TD: 6.3% (3) boat-ship: 4.1% (4) sailing: 3.0% (5) map: 2.7%
Occupy			(1) demonstr. (2) street (3) tent	(1) TD: 9.8% (2) VE: 7.8% (3) mccain: 7.3% (4) tony-blair: 6.9% (5) CTM: 6.3%
Champions League			(1) soccer (2) football (3) press-conf.	(1) VE: 14.5% (2) rugby: 14.4% (3) soccer: 14.4% (4) TD: 10.0% (5) CTM: 6.1%

the trend "ios5"), while others are reasonable (like "singing" for "Whitney Houston"). The last column displays the best systems for detecting the different events. Here, also the individual concept detectors are ranked, which indicates that some matched concepts are suitable for recognition (like "boxing" for "Mayweather-vs-Ortiz"). In general, for most trends either the (TD) or (VE) strategy ranks at the top for all evaluated concept detectors (with some exceptions for poorly recognized trends).

3.6 Discussion

This chapter presents a novel approach towards evolving vocabularies for video concept detection, which allows to provide a system being synchronized to current real-world events and therefore adapt to users' information needs. To accomplish such a synchronization multiple media channels are automatically mined to discover *trending topics* - subjects, which currently experience a high interest in social media.

In order to understand the dynamics in these channels, a comprehensive study was conducted on a large set of identified topics (n=2,986) covering the time period of an entire year. The analysis revealed that trending topics on Twitter and Wikipedia are more ephemeral than on Google, both rising and declining rapidly for newly emerging topics and that the observed media channels tend to specialize in specific topic categories.

Furthermore, to extend concept detection to evolving vocabularies it is critical for such systems to identify high-impact topics not only by today's momentum but more importantly by forecasting their life cycle as they emerge. Therefore, a fully automatic forecasting method for trending topics was proposed exploiting semantic similarity between topics. This approach – as evaluated on a large-scale dataset of Wikipedia viewing statistics – demonstrated superior forecasting performance (n=7,224) with a Mean Average Percentage Error of 45 % for a forecast of 14 days decreasing to 19 % for a 1-day forecast.

Finally, to demonstrate the capability of dynamic vocabularies for concept detection, different strategies for visual detection of trending topics in videos were presented. These strategies include a mapping of trending topics to large but static vocabulary of a concept detection system being trained on web video, a direct visual training of trending topics, and a combination of both, a dynamic extension of the static vocabulary with the trend detector. It could be seen in experiments on YouTube, that visual learning of these trending topics improves concept detection accuracy (n=65,000) by over 100% over static vocabularies and an additional marginal improvement could be achieved by the extension of static vocabularies (combination of both strategies). Interestingly, for some individual trending topics the support of a static vocabulary in combination with a trend detector made a big difference in detection accuracy (up to 5% in absolute AvgP). In addition, it was demonstrated that concept detection systems can be trained on demand utilizing a third party cloud computing infrastructure.

As future work, an interesting question is how concept vocabulary size might influence the detection accuracy of trending topics. It can be expected that an increased vocabulary size might have a positive influence of the proposed trend-to-vocabulary mapping. Also, a recent direction in concept detection is to understand which concepts from a large-scale vocabulary contribute the most information for the detection of more complex structures such as events [MGvdSS13a]. Following this idea, an investigation whether subsets of a fixed vocabulary would be of any help for a successful concept-to-trend mapping. Also, since the proposed marginalization method is utilizing trend to concept similarities, it might be of benefit to evaluate concept similarities according to knowledge bases such as WordNet based similarity

measures [Fel98] or DBPedia [ABK+07]. The potential of *topic based search* is also considered by Google, as they have extended their YouTube's API access mechanism towards topic[6].

[6] https://developers.google.com/youtube/v3/

Chapter 4

Training Data Retrieval and Active Relevance Filtering

To align concept detection with the latest user interest an efficient concept learning from the web is crucial. Unfortunately, current systems are troubled with the retrieval of relevant training material from platforms like YouTube and handicapped by the subjective and coarse nature of user-generated tags (or pseudo labels), which are only weak indicators of true concept presence. To remove these constraints, this chapter suggests an automatic *concept-to-query mapping* for high quality training data retrieval and *active relevance filtering* to generate high-quality annotations from web video tags. The key contributions of this chapter are[1]

1. An investigation of data retrieval from YouTube is presented, which quantifies the fraction of relevant content in web video as retrieved by LSCOM names (29%) and human refined queries (51%) (n=18,000).

2. A novel *concept-to-query mapping* method is introduced allowing to automatically retrieve relevant video material for concept training without the need of human query refinement.

3. It is shown that a direct use of web video tags degrades the performance of concept detection by a relative loss of up to 22% (n=100,000).

4. A novel approach called *active relevance filtering* is suggested, which combines automatic relevance filtering [Ulg09] with active learning methods to tackle the challenging task of eliminating correlated but non-relevant training content with a minimum of user interaction.

5. In experiments on YouTube data active relevance filtering was found to outperform both, purely automatic filtering and active learning approaches leading to a reduction of required label inspections by 75% as compared to an entirely expert annotated training dataset (n=100,000)

[1] This chapter is based on the authors' work in [UBB10, BUB10, BUB11a, BUB11b]

4.1 Introduction

As digital video has become an important source of information and entertainment to millions of users, databases grow larger [YOU13] and retrieval becomes a difficult challenge. This is particularly due to the semantic gap [SWSJ00], the discrepancy between low-level features of a video signal on the one hand and the viewer's high-level interpretation of the video on the other. To bridge this gap, concept detection has been proposed, which aims at automatically mining video collections for semantic concepts such as objects ("airplane"), scene types ("cityscape"), and activities taking place ("interview"). Concept detection has been studied intensively over the last years (for an overview, see Chapter 2) and is considered to be the key building block of various video search prototypes [CHL+07, NJW+09, SS10]. However, the effort associated with the manual acquisition of training samples for many concepts leads to a scalability problem. This has the consequence that the size of concept vocabularies remains limited and dynamic changes of users' information needs have to be neglected (as already outlined in Chapter 3.

This circumstance raises the question whether a manual acquisition of training material can be substituted with other information sources. One such source is *web video*, which is available at a large scale from portals like YouTube[2]. Web video content is usually enriched with user-generated tags, which indicate the presence of concepts in a clip. Utilizing this tag information as class labels, concept detection systems could automatically harvest training material from the web and thus perform a scalable and dynamic concept learning [KCK06, SS09, UKSB08, USKB10].

However, prior to detector training, systems utilizing web video must first retrieve relevant video clips by sending a query to the desired platform. Often, these queries are carefully constructed by a human operator [UBB10] to disambiguate content which is downloaded. An example can be seen in Figure 4.1 (middle): a straightforward mapping of the target *concept* "car racing" to the trivial *query* "car racing" may yield a training set containing non-relevant videos about car driver interviews or clips about remote controlled cars. Knowing this, a manual query refinement to "car racing tournament -rc -interview" and a restriction to the category "Autos & Vehicle" or "Sports" would reduces ambiguity and increase the amount of relevant content for detector training. Unfortunately, due to the time consuming process of such a manual query construction the idea of dynamic systems as presented in Chapter 3 is practically not feasible.

Furthermore, YouTube tags are coarse and therefore an unreliable indicator of concept presence. Following, it is challenging to utilize such tags as label information for supervised machine learning. An example is given in Figure 4.1 (right), which illustrates that not all YouTube videos tagged with "car racing" does in fact show the concept. This is due to several reasons: first, annotation behavior is subjective, and – though a concept may seem present to a specific user with certain knowledge and expectations – it may not be in general. Second, web video tags – which are usually given on a global scope – do not tell us *when* in a video the concept appears. Consequently, training sets acquired from web video portals are noisy and contain only a certain amount of truly relevant material. Concept detectors trained on such weakly labeled data must be expected to come with significant performance loss [KCK06, SS09]. This problem is also known in the literature as *label noise* [WN07, DUBW09, UBB10, TYH+09], *weak labels* [GY08, ATY09, LH10], or *pseudo labels* [WN08, HKC06, WJN09].

One straightforward strategy to overcome this problem would be to manually refine both, the query

[2] www.youtube.com

Figure 4.1: **Left:** To learn the concept "car racing" a query is formulated to retrieve training material from online portals like YouTube. **Middle:** Unfortunately, a simple query mapping of the concept name will deliver a wide range of related video clips which are not all suitable for visual learning of the concept. **Right:** Sample frames from YouTube clips tagged with "car racing". While some frames do show the concept (center), other content is *non-relevant*. This poses a challenge for concept detector training.

and the raw web-based training set to reduce or discard non-relevant content. While this has been demonstrated to improve the performance of the resulting concept detectors [SS09], it is very time-consuming and does not scale. To reduce manual annotation effort to some extent, *active learning* strategies have been proposed [AQ07a, AQ08]: instead of annotating the whole dataset, manual labels are only given for a subset of "most informative" samples. This has been demonstrated to achieve remarkable time savings when learning concepts from TV-based datasets [AQ07b]. In the context of web data filtering, however, active learning does not make optimal use of the given labels: these are employed to update the classifier, but remain neglected as a valuable clue for filtering noise in the training set. This raises the question if active learning can be extended for a better filtering of web-based training sets.

A second, alternative solution is to filter noisy material automatically [GY08, LSW08, USKB08b, WS08]. This approach has been referred to as *relevance learning* [LSW08] or *relevance filtering* [USKB08b]. Its core idea is to identify non-relevant content automatically based on its distribution in feature space and discard it during system training. However, such automatic relevance filtering systems do not reach the accuracy of a careful manual labeling. Therefore, it seems reasonable to assume that relevance filtering could benefit from a few manually provided labels i.e. how further improvements can be achieved with minimal human intervention.

The key contribution of this chapter is two-fold: First, this work presents an *concept-to-query mapping* for the automatic construction of YouTube queries such that a proper context for visual learning can be established for video download. Employing the presented query construction approach, it is demonstrated that the fraction of retrieved relevant content from YouTube is comparable to carefully human constructed queries. Second, a novel combination of the previous outlined approaches – active learning and relevance filtering – is suggested, which will be referred to as *active relevance filtering* in the following. This work proposes an interleaved setup of active learning label refinement and automatic relevance filtering. This way, the web-based training set is refined both manually and automatically during concept detector training. Using the proposed approach, it is demonstrated that concept detectors trained on weakly labeled web material can be improved significantly with a minimum of human supervision. Also, results show that the proposed active relevance filtering outperforms both a purely automatic noise removal and a standard manual refinement by active learning.

This chapter is organized as follows: first related work is discussed in the context of visual learning from web data (Section 4.2). Next, the concept-to-query mechanism is outlined in Section 4.3. After this, the proposed active relevance filtering framework is introduced (Section 4.4) and evaluated in quantitative experiments on web video data from YouTube (Section 4.5). A discussion concludes the chapter (Section 4.6).

4.2 Related Work

This section provides an overview of work related to training material acquisition for visual learning. As already outlined in Chapter 2, concept detection build upon supervised machine learning and requires a labeled dataset for classifier training. Such a dataset of positive and negative concept labels has to be either acquired from the ground or retrieved from the web. Unfortunately, user-generated tags from the web are only weak indicators of concept presence and therefore require additional treatment during detector training. The section starts with conventional label acquisition approaches such as active learning allowing to label a given dataset such that only a fraction of the data is inspected by a human operator. The section moves on with the usage of web video for concept detection. Here the focus is put on proper query construction for retrieval of training data and finally outlines approaches, which have been made in the context of dealing with pseudo labels associated with web video as a weak labeled data source with significant amount of label noise.

4.2.1 Label Acquisition with Active Learning

In supervised learning, classifier training is performed on a labeled dataset. Prior to training such a labeled dataset must manually acquired by annotating unlabeled samples. This is an expensive and cost-intensive effort. The main goal of active learning is to select only the "most informative" samples for manual annotation and therefore to minimize the effort of labeling new datasets [Set09]. In particular, active learning works in iterative cycles, where each cycle consist of three steps: first, a model training. Second, a "query sample" selection based on this model, and third, the manual annotation of the selected sample. This user feedback is then included in the next cycle of active learning leading to a successively improved classifier.

One particular family of active learning algorithms which is specifically suitable for retrieval is pool-based active learning. In pool-based active learning the learner has access to a pool of unlabeled data and can request the user to label a certain amount of instances in the pool to improve retrieval results. A straightforward method of selecting samples is *most relevant sampling* [SB90] as it is motivated by the idea of relevance feedback. In the context of text retrieval, Lewis and Gale [LG94] introduced *uncertainty sampling*, which is also known as the close-to-boundary criterion [SC00, TK02]. Another approach coming from the text-retrieval domain is based on the estimated error reduction strategy by Roy and McCallum [RM01] rather than utilizing heuristic approaches like aforementioned.

In the context of image retrieval, Tong and Chang [CTGC05, TC01] proposed a version space reduction approach for sample selection. Clustering-based approaches were presented in [NS04b, QSH$^+$06], where the structure of the underlying data distribution is utilized. This provides a good foundation for the cold start of active learning and improves performance by not labeling redundant samples belonging

to the same cluster. Active learning also finds application in video retrieval [SHDW05]. A large-scale evaluation of standard active learning sample selection methods can be found in [AQ07a], where close to ground truth detector performance could be achieved by labeling only 15% of the original TRECVID 2006 dataset. Particularly interesting is the performance improvement when taking temporal information into account for sample selection, i.e actively selecting neighborhood shots of already positive annotated shots. Ayache and Quenot also embedded active learning methods within the TRECVID collaborative annotation effort [AQ08]. A more general view of active learning for multimedia can be found in [CCHW05, HLY+06], where active learning is the method behind the feedback mechanism of the proposed retrieval system.

Summarizing, active learning can be employed in the practical situations, where only few annotations are given, active learning can help to efficiently identify and annotate the most "informative samples" to increase training data size such as done in TRECVID. This setup (which usually starts from very few reliable initial labels [AQ07a, AQ08, CCHW05]) differs from the one studied in this chapter, as this work focuses on a *refinement* of large but partially relevant training sets. Despite this difference, however, an application of active learning in the context of visual learning from the web seems promising as it can be used to verify relevant content and eliminate non-relevant ones. Beyond this, this work proposes a novel combination of active learning sample selection with an automatic relevance filtering, which will be demonstrated to lead to even more robust concept detectors at less manual annotation cost. In contrast to conventional active learning, the given setup starts from a noisy training set and uses active learning for a refinement.

4.2.2 Visual Learning from Web Labels

Though visual learning from web content is a challenging problem, this information source has been acknowledged as an attractive basis for training flexible and scalable visual recognition systems. Its exploitation is now an active area of research [YB05, SSTK08, SS09, USKB10, ATY09, RMJ+09, TAP+10, YT11, KLS13].

Unfortunately, such data as acquired via text-based image search engines or from portals like Flickr or YouTube contains a significant amount of non-relevant content for concept training. In case of Google Image Search, Fergus et al. [FFFPZ05] and Schroff et al. [SCZ07] reported a label precision between 18% and 77% for 7 object categories, and 39% over 18 categories respectively. Similarly, Li [LSW09] evaluated user tagging precision on a large-scale dataset of $20k$ manually inspected Filckr images of 20 different categories. The observation was, that although the precision varies among categories on average only 52% of the image tags were providing suitable material for detector training. Another analysis on Flickr image tags was performed by Setz [SS09] for 20 concept definition of the TRECVID 2008 benchmark, which concluded that 56% of Flickr images can be used for video concept learning. For web video as retrieved from YouTube, Ulges [Ulg09] reported tag labeling previsions around 38.6% averaged over 10 concepts. In case of YouTube video – however – one additional challenge related to tag precision is the *label resolution problem* [GY08] i.e. label information in videos may be coarse as YouTube users tag their video globally and without any further localization along the video stream.

Obviously the amount of non-relevant content depends on the proper formulation of the query to retrieve the required training material from the web. As seen in Ulges [USKB10] it is necessary to refine already constructed queries manually such that a proper context for visual learning can be established

(e.g. exclude music video from the "Beach Boy", when aiming to retrieve scenes of a beach). This circumstance was evaluated to improve label precision by up to 17.5% [Ulg09]. Therefore an alternative direction of getting more relevant training material is the seamlessly improvement of the initial retrieval by directly manipulating the query, an area related to concept-based query expansion or mapping [NHT+07, WLLZ07, YH07]. This mechanism reformulates a "seed" query with the intention to improve retrieval performance to provide the most relevant results given a user information demand. Traditional *query-to-keywords* reformulation as found in [YH07], can be split into "term reweighing" or "query expansion" methods. The first group re-weights individual terms of a query for use in the underlying vector space model, whereas the second group of methods adds additional terms to the query. This expansion can be either done by manual adjustment or relevance feedback given by users [Roc71, WJR06], pseudo-relevance feedback [CNL+04, XC96], or statistics about the entire collection [DDF+90, Fel98, KNC05].

Another type of approaches are the so called *query-to-concept* approaches [SW09]. This methods deal with the situation, where a concept vocabulary is given and must be matched against a user's search query. One straightforward way to perform this matching is to let the user select concepts by himself. However, choosing from a large concept vocabulary users have difficulties to select a proper set of concepts so that automatic concept selection mechanism (i.e. query prediction) come into account. Given an information need expressed in natural language, a first approach besides simple word spotting methods is the use of the vector space model to match a query against the semantic description of a concept [NZKC06, SWvG+06b]. Additionally, traditional query expansion approaches can be used to introduce additional query terms for concept selection. A comprehensive evaluation about different query expansion methods can be found in [NHT+07]. Here, the differentiation is made according to lexical approaches (synonyms, hypernyms from dictionaries like WordNet) and statistical approaches (local or global term frequencies and co-occurrences).

In this chapter, however, the main goal is not to satisfy a human users information need during retrieval but to retrieve *automatically* video material which is suitable for visual learning of concept detectors. Therefore – from this chapters point of view – the aim not to find the most relevant combination of concept detectors for a given query but to find a *proper* query formulation for a given concept definition.

One part of query construction for e.g. YouTube is the assignment of a category. Research in the area of web video categorization was performed based on visual and tag information [BHK+09], text and social information [XW09] and large-scale web crawling and search engine log data [TAP+10]. The proposed category assignment proposed in this chapter is similar to [TAP+10], where first tags are recommended and according to this information categories are assigned. However, the setting is different compared to the presented approach since the method selects tags and assign categories purely based on concept information and not the visual content of an uploaded video. Rather, this work employs external data sources for category assignment as done in [CCS+10].

4.2.3 Dealing with Label Noise

As seen in the previous section, web data can be considered as an attractive source for detector training but comes at a cost of a significant amount of non-relevant content. Even with retrieval done by carefully refining queries on YouTube [USKB10] or by randomly constructed images IDs on Flickr [LSW09] the problem of tags being ambiguous, subjective, or coarse leads to a label precision at most between 38.6% and 52%.

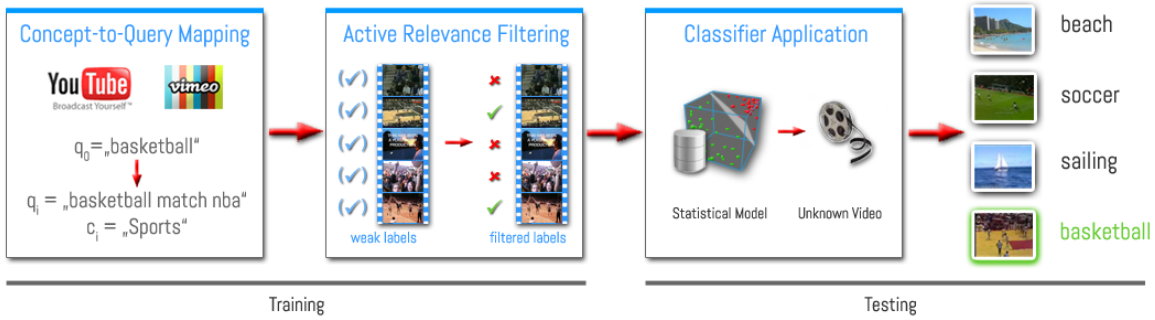


Figure 4.2: Concept learning from weakly labeled web video: material is downloaded from online platforms like YouTube, non-relevant content is filtered using *relevance filtering*, and a concept detector is trained, which can later be used to detect the learned concept in previously unseen videos.

Several approaches have been proposed to overcome this problem: one group of methods is targeted at a content-based refinement of raw web image sets [YB05, BF06, SCZ07, SSTK08]. This group of approaches start with a acquired image set from the web, identify a subset of "good" candidate images for concept presence either by manual inspection [BF06] or the analysis of surrounded image text [SCZ07, YB05]. These candidates are then taken to train a statistical model of the target concept and used to re-rank all remaining web images. Other methods closer to this work combine dataset refinement with model learning using topic models [FFFPZ05] or follow a semi-supervised learning approach like the OPTIMOL system [LWFF07] performing an iterative expansion of the training data. A third group of approaches perform relevance learning either by a nearest neighbor analysis of the data [LSW08, WS08, LSW10], where content is re-ranked by graph based random walks [HKC07, LHY+09], or identified to be filtered [WS08]. In context of weakly labeled video content, Gargi and Yagnik [GY08] emphasized the additional problem that label information in videos may be coarse, which they refer to as the *label resolution problem*. A bridge between image and video data was presented by [BBDB+10], where videos were enriched by image tag localized by visually similar Flickr images. Ulges et al. [USKB08b] presented a kernel-based approach for relevance filtering, such that the system automatically learns relevance weights during detector training.

The presented approach follows the relevance filtering line of research. Building upon the probabilistic setting presented in [USKB08b, WHS+06] it is extended by active learning to cope with the challenging setup where non-relevant content is correlated with the target concept in web videos. Overall, the usage of web images or video was demonstrated to be a valid alternative to expert annotated material for detector training. However, such content with its *pseudo labels* comes with a high amount of noise or non-relevant content. This affects concept learning by a significantly gradated detection performance [USKB08b, LSW09, SS09] and therefore is subject to current research efforts.

4.3 Concept-to-Query Mapping

In the following, a framework for query construction in the context of visual learning from the web is described. The system is outlined in Figure 4.2 (left box): to learn a concept like "basketball" video clips are retrieved from YouTube. This is realized by the construction of a query, which is send to

Table 4.1: **Automatic Keyword Selection**: Starting from a concept name, a set of keyword terms is selected based on synonyms retrieved from ImageNet, tag statistics and Google Sets items.

1. $q_0 = concept\ name$
2. retrieve synonyms s_{q_0} from ImageNet
3. retrieve tag statistic from YouTube
   - use q_0 to retrieve video data d_{q_0} from YouTube
   - calculate tag frequencies t_{q_0} from d_{q_0}
4. build $q_1 = \{s_0, ..., s_{n_s}\} \cup \{t_0, ..., t_{n_t}\} \subseteq s_{q_0} \cup t_{q_0}$
5. retrieve ranked list l_{q_1} from GoogleSets using q_1
6. build $q_i = \{t_0, ..., t_n\} \subseteq l_{q_1}$

the YouTube API. The core of the proposed approach is an automatic *keyword selection* and *category assignment* to a given concept. Taking the original LSCOM [KHN+06] concept name as initial query q_0, an automatic selection of keyword terms leads to an expanded query $q_i = \{baketball\ match\ nba\}$. This query is then used to infer a proper category $c_i = \{Autos\ \&\ Vehicles\}$ for the concept. Finally, the query q is constructed from q_i and c_i and used to retrieve training data for concept learning.

In the following, a query q represents the set of parameters which is used to retrieve videos from the YouTube API. This set of parameters may including text, tags, category restrictions and limitations to a particular time span or specific country. In this work however, the focus is given on the most distinctive parameter: keywords and categories, which lead to the query representation $q = \{t_0, ..., t_n\} * \{c_0, ..., c_m\}$ with n keywords t_i and m category assignments c_j. It should be kept in mind that the presented approach is general in the sense that it could be applied to all web video portals that allow access to their database through a similar API like YouTube.

4.3.1 Automatic Keyword Selection

The first step for the *concept-to-query mapping* is to transfer a concept name into a set of keywords. This procedure is illustrated in Table 4.1. The entry point is given by the concept name forming the initial query q_0. It is important to note that this initial query is expected to retrieve a significant amount of non-relevant videos. Based on q_0 a set of synonyms s_{q_0} is retrieved from ImageNet. Here, ImageNet is preferred over WordNet because it covers concepts suitable for visual learning. Additionally, tag statistics are calculated from the set of videos retrieved by q_0. For each tag appearing in this initial dataset, tag frequencies are calculated, which lead to a ranked list t_{q_0} of top tags for q_0. Although the process makes use of stop word removal and neglect digits and dates, this tag list can be considered noisy and less reliable than s_{q_0} for the purpose of disambiguation of concepts. However, t_{q_0} is intended to capture the specific wording of the YouTube community. By fusing s_{q_0} and t_{q_0} the new query q_1 is created: $q_1 = \{s_0, ..., s_{n_s}\} \cup \{t_0, ..., t_{n_t}\} \subseteq s_{q_0} \cup t_{q_0}$ with $n_s + n_t = n$ keywords. This query is now send to Google

Table 4.2: **Automatic Category Assignment**: Taking a query as imput, this procedure assigns a set of categories to the query using tag statistics and the hierarchical structure of the ImageNet Ontology.

1. use q_i to retrieve video data d_{q_i} from YouTube
2. calculate category distribution $p(c|d_{q_i})$
3. infer categories $c_{imagenet}$ from ImageNet
   - for each $t \in q_i$ get synset s_t from ImageNet
   - for each s_t get path p_t to ImageNet root
   - for each p_t get category c_t by $map(p_t)$
4. rank $c_{imagenet}$ according to $p(c|d_{q_i})$.
5. build $c_i = \{c_0, ..., c_m\} \subseteq c_{imagenet}$

Sets providing additional semantic relations in the context of q_1. Google Sets is a experimental prototype to generate lists of similar items. Its underlying probability model ranks these items according to their appearance in specific HTML structures as found in the world wide web. As a result a ranked list l_{q_1} of keywords is received, which after limiting it to n keywords results in the final query $q_i = \{t_0, ..., t_n\} \subseteq l_{q_1}$.

4.3.2 Automatic Category Assignment

As a second step the automatically assignment of categories for the previously constructed query q_i is presented. The procedure is outlined in Table 4.2. Given q_i, a second set of videos is retrieved from YouTube and its category distribution $p(c|videos)$ is calculated. Additionally, for each keyword $t \in q_i$ its corresponding ImageNet synset is found. If no synset is found for $t \in q_i$, this term will not contribute to the category assignment. For each synset found, the path from the synset node to the ImageNet root is build and mapped to a YouTube category according to a manual constructed mapping function $map(p)$. This mapping function maps ImageNet's first (and partially second) level synsets to YouTube categories allowing to transfer all $17k$ synsets to the given 15 YouTube categories by only providing roughly 60 manual mappings. A mapping in this context may just be as straightforward as $Animal \rightarrow Animals$ or as complex as $University \rightarrow Education$. The final step in the category assignment is a ranking of the mapped YouTube categories according to their query dependent distribution $p(c|videos)$ providing the set $c_i = \{c_0, ..., c_m\}$.

The final query q_n can now be constructed by $q_n = \{t_0, ..., t_n\} * \{c_0, ..., c_m\}$. This procedure provides an automatic query construction for training data retrieval from YouTube, which can be used for on-demand concept detector training as illustrated in Chaper 3.

4.4 Active Relevance Filtering

Although a carefully constructed query can help to disambiguate video retrieval for detector training, the coarseness of web video tags provides only a weak indicator of concept presence within video stream. In the following, a framework for visual concept learning from weakly labeled web video is described. The system is illustrated in Figure 4.2 (right box): to learn a concept like "basketball", training material is downloaded from online platforms. The core of the system – and the focus of this chapter – is a filtering of this weakly labeled web content, which identifies non-relevant material and performs a concept detector training in parallel. This process is referred to as *relevance filtering*, and is highlighted in a box in Figure 4.2. The procedure yields a statistical model (concept detector) which can then be applied to find the concept of interest in previously unseen video material.

Relevance filtering can be performed by one of the following three strategies (as seen in Figure 4.3):

1. an *automatic* relevance filtering, where non-relevant content is identified based on its distribution in feature space.

2. a *manual refinement* with the support of active learning, which selects the "most informative" samples for the user to label.

3. an *active relevance filtering*, which is the key contribution of this chapter and combines the two previous strategies by alternatingly performing automatic relevance filtering and a manual label refinement

In this section, first some basic notation are introduced (Section 4.4.1). After this, the two standard strategies will be addressed in detail, namely active learning (Section 4.4.2) and *automatic* relevance filtering (Section 4.4.3). Finally, the novel active relevance filtering approach is presented (Section 4.4.4).

4.4.1 Basic Concepts

In the following, video content is represented by keyframes, each associated with a feature vector $x \in \mathbb{R}^d$. For each concept of interest, a binary classification problem is formulated: the presence of the target concept is denoted with a label y, such that $y = 1$ indicates concept presence and $y = -1$ concept absence. The goal of concept detection is – given a keyframe x – to estimate the associated concept label y (or its probability $P(y = 1|x)$, respectively).

For training, a set of keyframes $x_1, ..., x_n$ is assumed to be given. Each of these is associated with a label $y_i \in \{-1, 1\}$ that indicates concept presence. In the setup of weakly labeled web videos, however, this true label is *latent* (i.e., not known), and only a weak indicator of concept presence is given (in practice, this is a tag given to the corresponding web video clip). This information is denoted by a *pseudo label* $\tilde{y}_i \in \{-1, 1\}$, and forms the input to the presented concept learning procedure.

It should be kept in mind that the approaches discussed in the following – particularly, the proposed active relevance filtering – could be applied as a wrapper around a variety of statistical models. In this chapter, this filtering approach is demonstrated for *kernel densities* as a well-known standard approach that has successfully been used for concept detection before [WHS+06, YSR05]

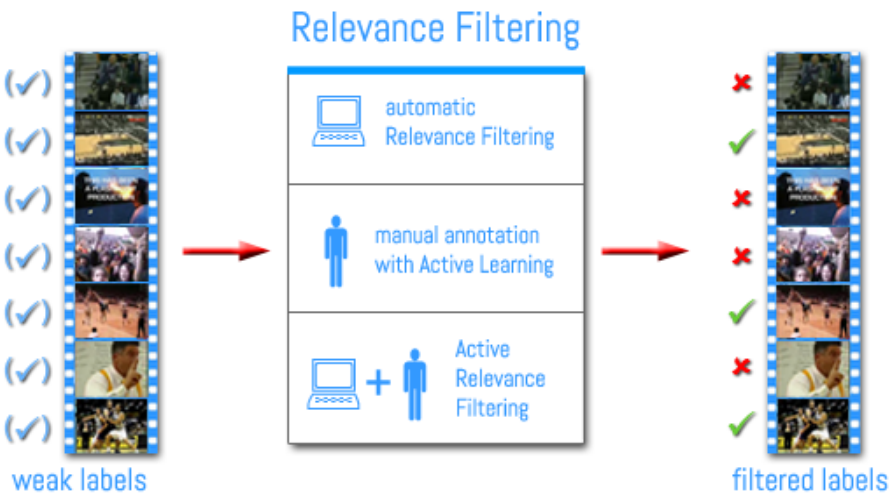


Figure 4.3: Relevance filtering, as illustrated here for the concept "basketball", can be performed using three strategies: (1) an *automatic relevance filtering* [USKB08b], (2) a manual refinement with the help of *active learning* [AQ07a], or (3) a novel interleaved combination of automatic and manual filtering called *active relevance filtering*, which is the key contribution of this chapter.

Kernel Density Estimation Baseline

First a simple supervised standard model is introduced that does not take label noise into account and will serve as a baseline in later experiments. This model uses two class-conditional distributions: p^1, which models positive keyframes (showing the target concept), and p^0 for negative frames (not showing the concept):

$$\begin{aligned} p^1(x) &= \frac{1}{Z_1} \cdot \sum_{i:\tilde{y}_i=1} K_h(x; x_i), \\ p^0(x) &= \frac{1}{Z_0} \cdot \sum_{i:\tilde{y}_i=-1} K_h(x; x_i). \end{aligned} \tag{4.1}$$

Z_1 and Z_0 are normalization factors. As a kernel function K_h, the well-known Epanechnikov kernel with Euclidean distance function and bandwidth h is used [DHS00, Ch. 4]:

$$K_h(x; x') = \frac{3}{4} \cdot \left(1 - \frac{||x - x'||^2}{h^2}\right) \cdot 1_{(||x-x'|| \leq h)} \tag{4.2}$$

By evaluating p^1 and p^0, the frame x is scored using Bayes' rule (the class prior is assumed to be uniform):

$$P(y = 1|x) = \frac{p^1(x)}{p^1(x) + p^0(x)} \tag{4.3}$$

It is important to note that the approach – as introduced so far – does not take the unreliability of web-based training labels $\tilde{y}_i$ into account. Instead, these labels are treated just like in a fully supervised

setup. Particularly, each positive sample ($\tilde{y}_i = 1$) though it does not necessarily show the concept (as illustrated in Figure 4.1) contributes to the density of positive samples p^1.

The key concern of this work, however, is to adapt concept training to the fact that user-generated labels on the web are inherently unreliable. In the following sections, several approaches for dealing with label weakness will be discussed. The basic assumption is that the given labels $\tilde{y}_i$ are only unreliable indicators of the true (but unknown) labels y_i such that:

- If the weak label is negative ($\tilde{y}_i = -1$), the true label is negative as well ($y_i = -1$).
- If the weak label is positive ($\tilde{y}_i = 1$), the sample *may* belong to the positive class, but does not necessarily do so, i.e. the true label y_i is unknown,

Briefly speaking, it is assumed that negative labels are reliable, but positive ones are not. This setup does not take false negatives ($\tilde{y}_i = -1$ and $y_i = 1$) into account, which is not strictly true (for example, a user could simply forget to tag a clip). According to observations on real-world web video, however, the fraction of these false negatives compared to truly negative content is negligible, and false positives pose a much more urgent problem.

4.4.2 Active Learning

One strategy to overcome label noise is to manually refine the raw web-based training set. In this context, active learning is a well-known effective approach. In this section, different active learning strategies for detector training are outlined that are targeted at achieving a label inspection at minimal additional annotation cost. The goal is to select only the most important samples for inspection and therefore to improve concept detector performance up to the level of ground truth expert labels with only a few manual labels.

Relevance Feedback as a Wrapper

In the following setup, a manual label refinement of selected samples is placed as a wrapper around a regular supervised learning method (the proposed kernel density learning from Equation (4.1)).

The procedure is illustrated in detail in Table 4.3: iteratively, concept detection is applied, obtaining class posterior probabilities $p^j = (p_1^j, ..., p_n^j)$ for all training samples, where $p_i^j \approx P(y_i = 1|x_i)$ (see Equation (4.3)). Based on these values, a keyframe $s^* \in \{i : \tilde{y}_i = 1\}$ is selected for manual annotation (here, the focus is given on *positive* weakly labeled keyframes because their labels are the unreliable ones). After a manual labeling of the selected sample s^*, its label is fixed to either -1 or 1 depending on the received annotation result. Note, that in case of a positive feedback (i.e., $\tilde{y}_{s^*} = 1$), no change of the model will occur, whereas in case of negative feedback, the associated label turns to be -1 and the model will change in the next iteration, resulting in an improved concept detector. This retrained concept detector will then provide new posterior probabilities for the next iteration of active learning sample selection. When continuing further, this procedure acquires more and more expert labels, until finally the weakly labeled dataset turns into a strongly annotated one.

Table 4.3: **Active Learning**: Wrapped around concept detector training, active learning selects informative samples for refinement by a user. Once the sample is labeled, its label is fixed to either -1 or 1 and the system is re-trained.

1. for $j = 1, ., m$ do:
   - obtain class posteriors $p^j = (p_1^j, ..., p_n^j)$ from p^1 and p^0
   - select sample s^* according to an *active learning* criterion Q:
   $$s^* := \underset{i:\tilde{y}_i=1}{\arg\max} \quad Q(p_i^j)$$
   - get the true label y_{s^*} from a human expert
   - fix the sample label:
   $$\tilde{y}_{s^*}^{j+1,...,m} = \begin{cases} 1, & y_{s^*} = 1 \\ -1, & y_{s^*} = -1 \end{cases}$$

   Once the true label is retrieved, the sample s^* is excluded from sample selection.

Active Learning Methods

Obviously, the quality of active learning heavily depends on the sample selection strategy Q (see Table 4.3). In the literature, many criteria Q have been proposed [Set09]. Here, the most popular ones are compared:

1. **random sampling**: samples are selected randomly (serves as a baseline).
2. **most relevant**: samples are selected which are most likely to be relevant and are therefore associated with the highest posterior [SB90]:
$$Q_{REL}(p_i^j) := p_i^j$$
3. **uncertainty**: samples are selected for which the relevance filtering method is least confident, i.e. $p_i^j \approx 0.5$ [LG94]:
$$Q_{UNC}(p_i^j) := 1 - |p_i^j - 0.5|$$
4. **density-weighted repulsion (DWR)**: This approach enhances "most relevant" sampling with an exploratory component. This is motivated by the assumption that the labels associated with clusters in feature space are homogeneous, and therefore the refinement of one sample within a cluster is sufficient of infer the remaining ones. This is realized by adding a repulsion term that enforces the query sample x_i to be distant from previously labeled samples (which form a kernel density p^+):
$$Q_{DWR}(p_i^j) := Q_{REL}(p_i^j) \cdot (p^+(x_i) + \epsilon)^{-\gamma},$$
where the parameter γ determines the strength of repulsion.

4.4.3 Automatic Relevance Filtering

While the active learning approaches introduced in the last section perform a refinement based on manual labels of selected samples, other systems have been introduced that replace this refinement with a fully automatic one. The basic idea of these *automatic relevance filtering* methods [USKB08b, WS08] is that relevant content appears frequently and forms clusters in feature space, while non-relevant material comes as outliers that can be identified and relabeled.

This section introduces an automatic relevance filtering approach based on a *weighted kernel density model* [USKB08b, WHS+06]. The class-conditional densities from Equation (4.1) are replaced with *weighted* kernel densities:

$$\begin{aligned} p^1_\beta(x) &= \frac{1}{Z'_1} \cdot \sum_{i=1}^{n} \beta_i \cdot K_h(x; x_i), \\ p^0_\beta(x) &= \frac{1}{Z'_0} \cdot \sum_{i=1}^{n} (1-\beta_i) \cdot K_h(x; x_i), \end{aligned} \tag{4.4}$$

where $Z'_1 = \sum_i \beta_i$ and $Z'_0 = n - Z'_1$ are normalization constants. Compared to the fully supervised setup from Equation (4.1), the key difference is that p^1 and p^0 are now parameterized by a vector $\beta = (\beta_1, ..., \beta_n)$. This vector consists of *relevance scores* $\beta_i := P(y_i = 1|\tilde{y}_i, x_i)$, meaning that each training sample is weighted according to its probability of being relevant: if a sample is likely to be relevant, it has a strong influence on the distribution of positive samples p^1_β but low influence on p^0_β. This way, the uncertainty of label information is taken into account.

To compute the class-conditional densities p^1_β and p^0_β, the vector of relevance scores β must be inferred in system training, i.e. potentially relevant frames must be divided into actually relevant ones and non-relevant ones.

The relevance scores β are estimated in a training procedure that – starting from a vector β^0 – iteratively updates the parameter vector β^k to a new version β^{k+1} by plugging it into the class-conditional densities $p^1_{\beta^k}$ and $p^0_{\beta^k}$ (Equation (4.4)). From these densities, new estimates of relevance scores can be obtained using Bayes' rule:

$$\begin{aligned} \beta_i^{k+1} &:= P(y_i = 1|x_i, \tilde{y}_i = 1) \\ &\approx \frac{P(y_i = 1|\tilde{y}_i = 1) \cdot p(x_i|y_i = 1)}{\sum_{y\in\{-1,1\}} P(y_i = y|\tilde{y}_i = 1) \cdot p(x_i|y_i = y)} \qquad (4.5) \\ &\approx \frac{\alpha \cdot p^1_{\beta^k}(x_i)}{\alpha \cdot p^1_{\beta^k}(x_i) + (1-\alpha) \cdot p^0_{\beta^k}(x_i)} \qquad (4.6) \end{aligned}$$

This is repeated until convergence. Training is regulated by the relevance fraction $\alpha := P(y_i = 1|\tilde{y}_i = 1)$, which determines how many of the positively labeled samples do in fact show the target concept (if $\alpha = 1$ is chosen, the model degenerates to the supervised case as in Equation (4.1)). In the following, it is assumed that a sufficiently good estimate of this parameter is be given.

Intuitively, this training procedure identifies regions in feature space where positively labeled frames concentrate and assigns high relevance scores to them, while outliers similar to negative content are

Table 4.4: **Active Relevance Filtering**: Wrapped around relevance filtering, active learning selects informative samples for refinement by a user. Once the sample is annotated, the system is re-trained and the remaining relevance scores are adapted.

1. for $j = 1, ..., m$ do:

   - **apply automatic relevance filtering, obtaining relevance scores** $\beta^j = (\beta_1^j, ..., \beta_n^j)$
   - **update the class-conditional densities** p_β^0 **and** p_β^1 **(Equation** (4.4)**)**
   - obtain class posteriors $p^j = (p_1^j, ..., p_n^j)$ from p_β^1 and p_β^0
   - select sample s^* according to an *active learning* criterion Q:

     $$s^* := \underset{i:\tilde{y}_i=1}{\arg\max} \quad Q(p_i^j)$$

   - get the true label y_{s^*}
   - fix the sample label:

     $$\tilde{y}_{s^*}^{j+1,...,m} = \begin{cases} 1, & y_{s^*} = 1 \\ -1, & y_{s^*} = -1 \end{cases}$$

     Once the label is fixed, its relevance score is set to the true value, and the sample is excluded from further automatic relevance filtering.

given low relevance scores. The approach resembles the well-known Expectation Maximization (EM) algorithm [DLR77], which maximizes the data likelihood in the presence of latent variables (here, the true concept labels $y_1, ..., y_n$). Also, a similar training procedure has been used by Wang et al. [WHS+06], but the system is constrained in a different way. While Wang et al. addressed a strictly semi-supervised setup – where initial reliable training samples for all classes are available — this work cannot rely on such information in the given weakly supervised setup. Instead, this method constraints the system with a certain *prior* of expected relevant material α. For more information on the approach, please refer to a previous publication [USKB08b].

4.4.4 Active Relevance Filtering

The relevance scores $\beta_1, ..., \beta_n$ in Section 4.4.3 captured the uncertainty of the given web-based label information. They have been fitted using an automatic training procedure, which has previously been shown to improve concept detection to some extent [USKB08b]. Yet, significant label uncertainty remains, which is why a combination of relevance filtering with active learning is proposed to enhance the system with a limited amount of manual feedback. This *active relevance filtering* is outlined in the following.

Enhancing the Relevance Feedback Wrapper

Next an iterative manual labeling of selected frames is suggested, which is alternated with a retraining of relevance scores β. This way, the previously introduced active learning mechanism is enhanced by an automatic relevance filtering step after concept detector training. Again, to reduce annotation effort, active learning strategies are used to select only the most informative samples for annotation.

The procedure is illustrated in Table 4.4 (modifications compared to the active learning procedure in Table 4.3 are highlighted in bold): in each iteration, an automatic relevance filtering is performed, from which the class-conditional densities are updated, obtaining class posteriors $p_1^j, ..., p_n^j$ for all training samples. Based on these posteriors, the most informative weakly labeled keyframes are selected for manual annotation (the same selection strategies as for active learning can be used, see Section 4.4.2). The received label information will now again serve as additional ground truth for the next iteration of automatic relevance filtering, providing improved relevance scores for the next iteration of sample selection. With more iterations of such combined relevance filtering and active learning, the procedure separates relevant content from non-relevant one more reliably.

Note that this approach *alternates* automatic and manual filtering: in contrast to a purely automatic filtering, the method uses an additional wrapper in which a human operator contributes more accurate labels than the purely automatic approach can estimate by itself. The key difference to active learning is that the labels are not only used to update the classifier, but also for further relevance filtering: each time a new label is given, it influences relevance scores on the training set and helps to filter non-relevant content more precisely. This provides an improved basis for the next active learning sample selection – alternatingly, automatic and manual refinement boost each other.

4.5 Experimental Evaluation

Experiments are performed on two separate datasets of web video content downloaded from YouTube. The first dataset serves as foundation for the evaluation of the proposed concept-to-query mapping. The second dataset is already known in the literature from previous evaluations of relevance filtering [USKB08b, Ulg09] and is therefore used to test the suggested active relevance filtering approach.

Four experiments were conducted to quantify the effects of different refinement strategies: First, the automatic query construction approach is evaluated against a careful query construction by a human operator (Section 4.5.1). Second, the impact of an automatic relevance filtering is validated to demonstrate that this approach gives some improvements but does not reach the performance of a complete manual annotation (Section 4.5.2). After this, a manual refinement using plain active learning is evaluated (Section 4.5.3) and compared with the novel active relevance filtering approach (Section 4.5.4).

4.5.1 Comcept-to-query Mapping

To evaluate the proposed automatic mechanism of query construction 30 concepts from the TRECVID 2011 benchmark, which have been selected by NIST for evaluation have been taken. They include concepts related to objects (“flowers”, “boat-ship”), locations (“cityscape”, “mountain), or sports (“swimming”, car racing”). For each concept three experiments have been performed:

Table 4.5: Results of the concept-to-query mapping evaluation. Fractions of relevant material i.e fraction of videos containing the concept are displayed for each concept and each of the three experiments. The last line displays the relevance as average over all 30 concepts.

Concept Name	**[exp-1]**	**[exp-2]**	**[exp-3]**	**[exp-3] Queries (keywords; category)**
airplane flying	0.30	0.47	**0.56**	airplane flying aircraft; Autos/Vehicles
animal	0.40	0.89	**0.93**	animal nature; Pets/Animals
Asian people	0.36	**0.52**	0.40	asian people asians; People/Blog
bicycling	0.29	**0.63**	0.62	bicycling city; Sport
boat ship	0.30	0.57	**0.68**	boat ship water; Autos/Vehicles
bus	0.15	0.57	**0.72**	bus buses; Autos/Vehicle
car racing	0.48	0.50	**0.64**	car racing cars; Autos/Vehicles
cheering	0.48	0.27	**0.54**	cheering cheer; Sports
cityscape	0.13	0.12	**0.14**	cityscape architecture; Travel/Events
classroom	0.13	**0.39**	0.36	classroom students; Education
dancing	0.53	0.56	**0.61**	dancing live; None
dark-skinned people	0.62	0.65	**0.79**	dark skinned people; People/Blogs
demonstration or protest	**0.76**	0.72	0.41	demonstration protest funny; News/Politics
doorway	0.07	**0.20**	0.15	doorway vent; Howto/Style
explosion fire	0.33	0.35	**0.61**	explosion fire gasoline; How/Style
female human face closeup	0.04	**0.64**	0.36	female human face closeup; None
flowers	0.11	**0.53**	0.42	flowers green; Howto/Style
ground vehicle	0.31	0.47	**0.70**	ground vehicle military; Autos/Vehicles
hand	0.19	0.51	**0.59**	hand; Science/Technology
mountain	0.11	0.61	**0.70**	mountain peak; Travel/Events
nighttime	0.05	0.28	**0.53**	nighttime building; Travel/Events
old people	0.23	0.23	**0.36**	old people; Comedy
running	0.28	0.35	**0.63**	running basketball; Sports
singing	0.78	**0.88**	0.63	singing fun; None
sitting down	0.02	**0.10**	0.06	sitting down the; Travel/Events
swimming	0.48	**0.78**	0.70	swimming water; Sport
telephones	0.04	**0.67**	0.46	telephone call; Science/Technology
throwing	0.13	**0.65**	0.16	throwing to;None
vehicle	0.31	**0.64**	**0.64**	vehicle car; Autos/Vehicles
walking	0.20	**0.44**	0.09	walking alternative; None
average	0.29	**0.51**	**0.51**	

- **[exp-1]** query construction by a simple one-to-one mapping of concept name to a YouTube query. Here, concept names from LSCOM are taken.
- **[exp-2]** query construction by manual refinement from a human according to a visual inspection on YouTube.
- **[exp-3]** query construction performed by the proposed automatic concept-to-query mapping from Section 4.3. Queries were limited to $n = 3$ keywords and $m = 1$ category assignments.

While standard training sets for supervised learning do provide accurate label information, positive training samples in web video datasets contain only a particular fraction of relevant samples. This *relevance fraction* is denoted with α in the following. In case of web video α can be considered as measure of label noise. Consequently, for expert labeled datasets it can be expected to observe an α close to 1.0 whereas for web video an α significantly lower that 1.0 is presumed. Obviously, the value of α may differ among concepts. To provide an first insights of α for web video downloads, for each query and experimental setup, 100 videos are downloaded from YouTube and manually reviewed according their

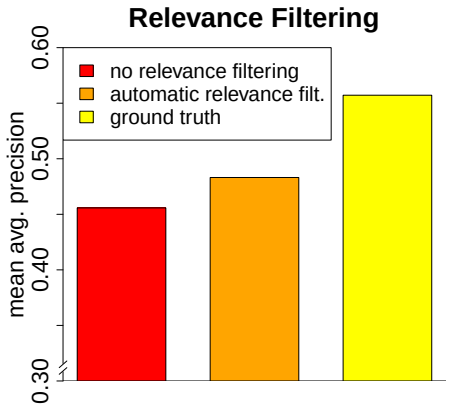


Figure 4.4: Results of the baseline experiment, showing potential performance ranges for further refinement strategies. Though automatic relevance filtering provides some performance gain, its performance is far from the ground truth optimum. Note that ground truth label information is not given and should here only demonstrate the potential performance gain of a perfect relevance filtering.

relevance (i.e.,α) to the LSCOM concept definition. This manual inspection which is based on three keyframes per video clip evaluates how many of the retrieved video clips truly contain the concept.

Table 4.5 illustrates the results of the evaluation. For each concept the fraction $\alpha \leq 1.0$ of relevant content is shown and additionally for **[exp-3]** the automatically constructed queries are printed. When comparing these three experiments, it can be seen that **[exp-1]** queries perform weak i.e. they contain the most non-relevant content when retrieving videos from YouTube. Further, manual refined queries **[exp-2]** and automatically constructed queries **[exp-3]** perform comparable to each other improving the fraction of relevant content by 76%.

For some concepts like "airplane flying" or "boat ship" the approach particularly benefits from ImageNet synonyms whereas for concepts, where no synonyms could be found the focus on frequent YouTube tags may lead query construction into the wrong direction like observed for the concept "throwing" or "singing". Also, for concept with a uncommon concept name like "female human face closeup" the method was not able to retrieve any content from YouTube. However, for the majority of concepts the selected keyword terms were semantically meaningful and related to the concept. Also, for most category assignments the method selected the same category a human operator would do.

4.5.2 Weak Label Impact & Automatic Relevance Filtering

To evaluate active relevance filtering, ten test concepts from the YouTube-22concepts [Ulg09] dataset are selected, including objects ("cats", "eiffeltower"), locations ("beach", "desert"), or sports ("basketball", "golf"). For each concept, 100 video clips were downloaded by querying the YouTube API with an appropriate combination of keywords. Keyframes were extracted and manually assessed according to canonical concept definitions. For each concept, a training set of $1,000$ negative sample frames and 500 noisy positive frames is sampled. The label precision of these positive samples was set to 20% (which was validated to be a typical value for web video in previous annotation experiments). This means that the 500 positive samples contained only 100 true positives and 400 false positives (which were also sampled

Table 4.6: Detailed results of the baseline experiments. Average precision is displayed for each concept and each of the three runs.

concept	**no relevance filtering**	**auto. relevance filtering**	**ground truth**
basketball	0.570	0.606	**0.651**
beach	0.398	0.449	**0.504**
cats	0.320	0.333	**0.388**
desert	0.587	0.636	**0.655**
eiffeltower	0.425	0.421	**0.526**
helicopter	0.362	0.392	**0.418**
sailing	0.440	0.466	**0.493**
soccer	0.562	0.575	**0.740**
swimming	0.448	0.491	**0.647**
tank	0.441	0.457	**0.543**
MAP	0.455	0.482	**0.557**

from YouTube clips tagged with the target concept, but were manually assessed to be non-relevant). To evaluate the concept detectors trained on this weakly labeled content, a test set of 500 positive and $1,500$ negative frames was sampled (it was made sure that training and test content was drawn from different clips).

As a feature representation of keyframes, this work refer to the well-known bag-of-visual-words approach [SZ06, vdSGS08a]: a regular patch sampling was conducted at several scales, patches were described by SIFT [Low99], and finally clustered to a $2,000$-dimensional vocabulary using K-Means. After this, a PLSA dimensionality reduction [QMO$^+$07] to 64 dimensions was applied for efficiency purposes. The relevance filtering system was tested with a kernel bandwidth of $h = 0.275$ (which was previously optimized using cross-validation). The parameter α of automatic relevance filtering (Equation (4.5)) was set to 20%. For DWR sampling a value of $\gamma = 0.1$ proved to work best. As a performance measure, *mean average precision* (MAP) is used. All results are averaged over all 10 test concepts and over 5 trials using different randomly sampled datasets.

This experiment, evaluates several concept learning approaches when trained on weakly labeled web video material. In total three systems are compared: first, one that does not perform relevance filtering at all, which corresponds to a standard supervised system using plain kernel densities (this baseline is denoted with *no relevance filtering* and has been outlined in Section 4.4.1). Second, an automatic relevance filtering as outlined in Section 4.4.3, and third a control run using ground truth labels (note that such label information is not available in practice).

When comparing these three runs (Figure 4.4 and Table 4.6 for detailed concept-dependent results), it can be seen that the system without relevance filtering performs worst, with a mean average precision (MAP) of 0.455. The automatic relevance filtering achieves a slight improvement (MAP: 0.482). However, a strong gap of 7% remains compared to the ground truth run (MAP: 0.557) – this indicates that concept learning from the web could be improved significantly if a more accurate filtering of non-relevant content can be performed. This motivates semi-automatic refinement strategies as evaluated in the next sections.

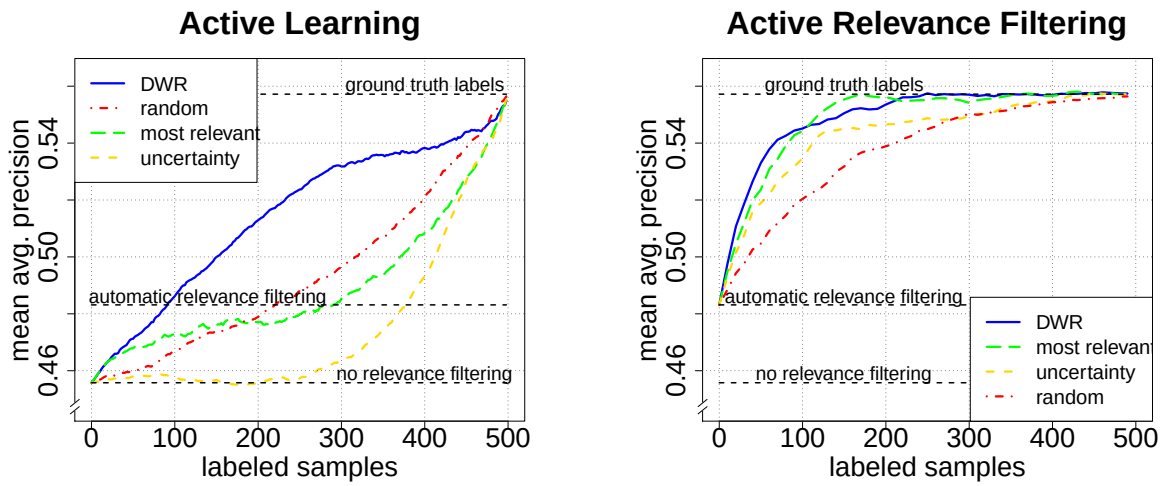


Figure 4.5: **Left:** Results of active learning. The accuracy of the resulting concept detectors is plotted against the number of manually annotated training samples. **Right:** Results of active relevance filtering. Performance is plotted against the number of manually annotated training samples. It can be seen that – if using a proper sample selection – it is sufficient to annotate only $30-40$ weakly positive training samples to achieve a significant performance improvement.

4.5.3 Active Learning

The third experiment quantifies the performance of a manual refinement of web-based training sets using active learning. The results of this experiment are illustrated in Figure 4.5 (left), where the performance of the trained concept detectors on the test set is plotted against the number of training samples annotated with active learning (different curves correspond to different sample selection strategies). To establish a relation to the last experiment, the three automatic runs (no relevance filtering, automatic relevance filtering, and ground truth labels) are plotted as dotted lines in Figure 4.5 (left). It can be seen that all sample selection methods start at an MAP of 0.45 (which equals the previously shown "no relevance filtering" system, as no automatic relevance filtering is done. However, as more training labels are collected manually, the quality of the training set (and with it the accuracy of the resulting detectors) improves. Sample selection stops when all weakly labeled samples are manually refinement i.e. after 500 annotations. Here, the MAP is the same for all selection methods and equals the "ground truth" run in Section 4.5.2 (which is not surprising, as the whole training set is now manually annotated).

When comparing the different sample selection methods, it can be seen that different sampling strategies lead to a very different performance. Surprisingly, well-known samplings methods like *uncertainty sampling* are performing worse than a simple random sampling baseline. The best overall result is achieved by DWR sampling, which gives strong improvements over all other strategies. Yet, the improvements by active learning remain limited: even the best method requires a substantial amount of manual samples to give significant improvements over the automatic relevance filtering. To reach a performance close to a ground truth labeling, all methods require a manual annotation of wide parts of the training set.

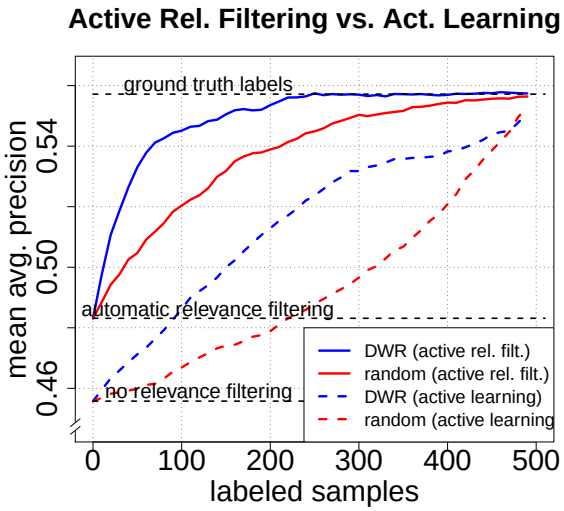


Figure 4.6: Comparing active learning and active relevance filtering, using random sample selection and DWR. The proposed active relevance filtering leads to better concept detectors at lower annotation cost.

4.5.4 Active Relevance Filtering

In this experiment, the performance of the proposed active relevance filtering approach (a novel combination of a manual and automatic label refinement) is evaluated. Results of this experiment are illustrated in Figure 4.5 (right). Just like in Figure 4.5 (left), concept detector performance is plotted against the number of manual annotations used in training.

First the different active learning strategies are compared in Figure 4.5 (right). It can be seen that all used sample selection methods outperform the random sampling baseline significantly. Systems based on *most relevant sampling* perform best, which can be explained by the fact that this approach helps to identify false positives that are "surprising" to the system and thus lead to strong model changes. For low numbers of annotations, the exploratory component of DWR leads to further improvements.

Overall, it can be seen that active relevance filtering — if combined with the right sample selection strategy — is highly efficient, giving strong improvements of concept learning even for very low numbers of manual annotations. For example, with as few as 50 annotations, a performance increase of MAP 5% is achieved compared to automatic relevance filtering. When continuing with annotation, it can be observed that concept detection performance converges to the ground truth case at $125 - 150$ iterations (which corresponds to only $25 - 30\%$ of the positive weakly labeled training set and 10% of the whole training set).

Figure 4.7 provides a visual impression of active relevance filtering performance. Here, the top 20 test set classification results are shown for the three concepts "basketball", "tank" and "eiffeltower". For each concept a separate result list is displayed for a) non relevance filtering, b) automatic relevance filtering and c) active relevance filtering (50th iteration of DWR). The border of each keyframe is colored according to its true label (green=concept present; red=concept absent). Comparing the different lists, a significantly better results can be achieved for c) compared to b), which itself shows improvements over a). Note that particularly for such challenging concepts as "eiffeltower", where automatic relevance

Figure 4.7: Results for the concepts "basketball" (top), "tank" (center), and "eiffeltower" (bottom). Top detections are displayed for a) no relevance filtering, b) automatic relevance filtering and c) active relevance filtering (DWR). Keyframes with green borders indicate a correct detection of the concept and red border a incorrect ones.

filtering struggles with, active relevance filtering improves classification results significantly.

Finally, Figure 4.6 compares the proposed active relevance filtering with pure active learning as discussed in Section 4.5.3. Again, detection performance is plotted against the number of manual annotations. The figure plots the best systems for each, active relevance filtering and active learning (Section 4.5.3 vs. Section 4.5.4), namely the DWR-based runs, and as a baseline random sampling. The results clearly indicate that active relevance filtering significantly outperforms a pure active learning. As seen, for both sample selection strategies that active relevance filtering starts with a higher MAP as it utilizes automatic relevance filtering. Also, system performance of active relevance filtering improves quicker than for pure active learning, which can be explained by the fact that active relevance filtering makes better use of user feedback: if a manual sample is provided, the additional relevance filtering mechanism propagates this label over neighbor samples. For example, after refining only 50 samples manually, active relevance filtering with DWR sampling (MAP: 0.54%) clearly outperforms pure active learning, with DWR (MAP: 0.47), resulting in an absolute improvement of 7%.

Concluding, active relevance filtering, particularly if combined with appropriate sample selection strategies, can improve concept learning on the difficult domain of web video content better than both an automatic relevance filtering and a manual label refinement using standard active learning techniques.

4.6 Discussion

In this chapter, the challenge of learning visual concepts from web video was addressed, which offers a scalable alternative to the conventional manual acquisition of concept detection training data. On the downside, the tags coming with web video are only weak indicators of concept presence, and web-based training sets come with significant amounts of non-relevant content. To achieve robustness with respect to such *label noise*, this work presents two contributions: First, it presents an automatically mapping of concepts to queries for training data retrieval is suggested. Second, it combines *relevance filtering*, which discards non-relevant content automatically and active learning, which is targeted at an efficient manual refinement. The resulting approach – called *active relevance filtering* – performs a highly efficient learning using a few manually labeled samples.

In quantitative experiments with real-world web content downloaded from YouTube, it has been demonstrated that (i) an automatic construction of queries for training data retrieval is achieved yielding the same high quality as done by humans and (ii) the proposed active relevance filtering approach improves concept learning significantly. Particularly, the proposed approach outperforms both a purely automatic refinement and standard active learning, reaching a performance comparable to ground truth training by refining only $25 - 30\%$ of weak positive labels in the training set.

Regarding future directions along this line of research, the next step is to integrate active relevance filtering with other statistical learning methods. In this chapter, a less complex generative standard approach was used (namely, kernel densities). It remains to be investigated whether active relevance filtering could be used as a wrapper around other machine learning methods in a similar fashion, including generative ones (e.g., Gaussian mixture models, histograms) as well as discriminative ones (e.g., SVMs [SS01]).

Chapter 5

Adjective Noun Pairs for Visual Sentiment Analysis

In this chapter, the challenge of sentiment analysis (i.e. the analysis of predictive judgments in the context of their polarity) from visual content is addressed. To achieve this goal, concept detection is extended by *Adjective Noun Pair* combinations, which serve as a novel mid-level presentation of images and videos. It is shown that the presented approach not only allows to capture the sentiment conveyed by images but also is able to detect emotion being reflected in images, and provide support the analysis of child sexual abuse (CSA) material. The key contributions of this chapter are[1]:

1. The first concept detection approach providing sentiment prediction from visual content.
2. The Visual Sentiment Ontology (VSO), a large-scale ontology of 3,000 Adjective Noun Pairs (ANP), which is founded on psychology theory and is constructed by a fully automatic data-driven methodology mining the web.
3. SentiBank, a novel mid-level representation framework, which builds upon the VSO and encodes concept presence of 1,200 ANPs. This detection bank reached an F-Score performance for ANP detections of 0.6 (n=500,000) and was made publicly available to foster research in this area.
4. The first public visual sentiment benchmark dataset consisting of 2,000 photo tweets and 19k crowd-sourced ground truth annotations
5. In three independent experiments on sentiment prediction (n=2,000), emotion detection (n=807) and pornographic filtering (n=40,000) the performance of ANPs represented by SentiBank was demonstrated to either outperform other low-level feature representations (sentiment prediction, pornography detection) or perform comparable to state-of-the art methods (emotion detection). As an addition, SentiBank allows for the explanation of system results by listing detected ANPs as visual ingredients of its detection.

Concluding, this effort – as being the first of its kind – created a large publicly available resource for for further investigation of Adjective Nouns Pairs as a novel extension for concept detection.

[1]This chapter is based on the authors' work in [BJC+13, BJCC13, SHBD14]

Figure 5.1: Tweets from the "2012 Year on Twitter" collection: Barack Obamas famous reelection tweet (left) and a tweet capturing the destruction caused by Hurricane Sandy (right). Both tweets display a generic text ("four more years" and "rollercoaster at sea" respectively) and convey the main information, including its sentiment, visually.

5.1 Introduction

Nowadays the Internet – as a major platform for communication and information exchange – provides a rich repository of people's opinion and sentiment states[2] on a vast spectrum of various topics. This knowledge is embedded in multiple facets, such as comments, tags, browsing actions, as well as image and video content. The analysis of such information either in the area of opinion mining, affective computing or sentiment analysis plays an important role in behavior sciences aiming to understand and predict human decision making [PL08] and enables applications such as brand monitoring [JZSC09], stock market prediction [BMZ11], political voting forecasts [OBRS10, TSSW10] or intelligence gathering [YN07].

So far, the computational analysis of sentiment concentrates on textual content [PL08]. Limited efforts are devoted to analyzing sentiments from visual content such as images and videos, which is becoming a pervasive media type on the web. For example, two of the most popular tweets in the year 2012 (see Figure 5.1) convey sentiment information primarily by visual means. Thus, an open issue with state-of-the-art sentiment analysis is the need of visual content analysis.

This problem poses a set of unique challenges as it addresses abstract human concepts in the sense of emotion and affect. Typically, semantic concept detection builds on the physical presence of objects or scenes like "car" or "building" being visible in an image. Sentiment could differ among persons as the stimuli evoked human responses are naturally subjective. An analogy is given by Machajdik [MH10] by formulating the *affective gap*[3] as counterpart to the *semantic gap* in CBIR. To fill the semantic gap, mid-level representations based on visual concepts have been proposed. In this work a similar proposal is made by the discovery and detection of a set of visual concepts that can be used to fill the affective gap and automatically infer the sentiment reflected in visual content. Please note that the presented mid-level representation is more expressive than the ones stated in [WH08] as it has capability to explain sentiment prediction results beyond color schemes.

In this chapter a novel approach towards sentiment analysis is presented, which is founded on the semantic understanding of visual content. For this purpose based on Plutchik's Wheel of Emotions [Plu80]

[2]Please note that, throughout this chapter sentiment will be defined similarly to [PL08], as the polarity of an opinion item which either can be *positive, neutral or negative*

[3]gap between low-level features and the emotional content of an image reflecting a particular sentiment

(a) Beautiful Clouds

(b) Dark Clouds

(c) Cute Dogs

(d) Dangerous Dogs

Figure 5.2: Four Adjective Noun Pair samples illustrating the capabilities of adjectives to change the sentiment an image conveys and the potential to separate the visual space representing a noun such as "dog".

a large-scale *Visual Sentiment Ontology (VSO)* of 3,000 semantic concepts is automatically constructed, with each concept being selected according to the following criteria: *(1)* reflect a strong sentiment, *(2)* has a link to an emotion, *(3)* be frequently used and *(4)* has reasonable detection accuracy.

To satisfy the above conditions, the idea of a semantic concept is extended to *Adjective Noun Pairs (ANP)* such as "beautiful flower" or "disgusting food". As seen in Figure 5.2 the advantage of ANPs, is their capability to turn a neutral noun like "clouds" or "dog" into an ANP with strong sentiment, like "beautiful clouds" or "cute dog" by adding an adjective with a strong positive sentiment and vice versa by adding a negative adjective an ANP can be turned into a strong sentiment one like "dark clouds" or "dangerous dogs". Such combined phrases also make the concepts more detectable, compared to adjectives only, e.g. an adjective concept like "beautiful" is abstract and hard to detect. Please note that by adding adjectives to nouns the visual context of such combined pair concepts changes significantly allowing to partition the corresponding visual space of nouns along the set of different adjective combinations. This brings unique opportunities to the construction of underlining ontologies for visual learning.

Building upon the VSO this work introduces *SentiBank*, a library of trained concept detectors providing a mid-level visual representation with respect to VSO criteria. It is shown - through extensive experiments - that reasonably reliable detector performance can be achieved for more than 1,200 ANP concepts, which form SentiBank. Further, experiments within different application domains demonstrate the usability of the proposed approach towards sentiment prediction: On image tweets, it improves state-of-the-art text-based prediction accuracy by an absolute gain of 13%, emotion detection with comparable results to state-of-the art methods, and CSA filtering with a reduction of EER from 14.0% to 8.3%. In summary, this work presents the first visual analysis approach for sentiment prediction known to the lit-

erature including - *(i)* - a systematic, data-driven methodology to construct an ontology from established folksonomies, - *(ii)* - the large-scale Visual Sentiment Ontology founded on a well-known psychological model, - *(iii)* - a mid-level representation built over this ontology helping to bridge the *affective gap* and - *(iv)* - the public release of the VSO and its large-scale dataset, the SentiBank detector library, and the benchmark for visual sentiment analysis.

In this chapter, first related work (Section 5.2) is discussed and the framework overview is given (Section 5.3). After this, the design and construction methodology of the VSO (Section 5.4) and its analysis (Section 5.5) are outlined. Further, SentiBank, the proposed mid-level attribute representation will be described (Section 5.6). Finally, several application of SentiBank will be shown and evaluated (Section 5.7) and the chapter closes with a discussion (Section 5.8).

5.2 Related Work

In this section an overview of research related to sentiment analysis is given. A review of conventional concept detection approaches is omitted here as is it already given in Chapter 2. Instead, the related work with respect to visual learning, which have a direct link to the presented work is covered. Starting with an outline of textual sentiment analysis dealing with the extraction of personal opinion from natural language, the section continues with the review of visual sentiment or related areas such as affect or emotion detection from visual content. In addition, work in the context to ontology construction and visual learning of mid-level feature or attribute representations will conclude the overview of related work.

5.2.1 Textual Sentiment Analysis

Research in the area of automatic text analysis of opinion and sentiment dates back at the beginning of the rise of the Internet [WWB01, DC01]. Originating from the field of subjectivity analysis [WR88] the terms “sentiment analysis” and “opinion mining” were first introduced in the beginning of the 21st century [NY03, DLP03]. According to the terminology proposed by [WWC05], the field focuses on the automatic identification of personal states (i.e. opinions, sentiments) in natural language.

The prediction of sentiment has a rich background in creating dictionaries of positive or negative words [ES06, WWH05] and the explicit investigation of polarities between words [ES06, TBP$^+$10]. Here, different approaches have been presented, ranging from rule-based systems such as in OpinionFinder [WR05], semi-supervised learning [ES06], or unsupervised bootstrapping approaches [Tur02], where initialized by two words such as “excellent” and “poor” a larger vocabulary is built by measuring semantic closeness of other words and phrases to these points at the positive-negative scale. A comparison between systems employing Naive Bayes, Maximum Entropy Classification, and SVMs for sentiment analysis can be found in [PLV02]. For further details please refer to the survey providing an overview in this area [PL08].

The work in this chapter is similar to the above as it also builds a lexicon of words with a strong sentiment. However, this lexicon is constructed particularly to serve for the visual learning of semantic concepts. For this purpose a careful selection of concepts has to be undertaken with respect to ontology construction in concept detection since not every word is visually learnable or detectable. For example,

the word "love" has a very strong positive sentiment but is not visually graspable. Nevertheless, the presented approach shares some similarity with the retrieval methods in [Tur02], who began to create a larger lexicon from only two strong opposite sentiment. The ontology construction process follows this idea and constructs the entire ontology automatically from 24 emotions serving as a seed vocabulary for the presented data-driven sentiment word discovery.

5.2.2 Visual Sentiment Analysis

As previously outlined, with respect to sentiment analysis much progress has been made on text analysis [ES06, TBP+10] and textual dictionary creation [ES06, WWH05]. However, efforts for visual analysis fall far behind. The closest that comes to sentiment analysis for visual content is the analysis of aesthetics [DJLW06, MPLC11], interestingness [IXTO11], and affect or emotions [JWW+12, MH10, YvGR+08, YUB+12] of images or web pages [WCLH11]. To this end, either low-level features are directly taken to predict emotion [LFXH12, JWW+12], or indirectly by facial expression detection [VW12], or user intent [HKL12]. Similarly Wang [WJH+12], who introduced a so called *high-level representation* of emotions, is limited indeed grouping low-level color scheme features. For more details please refer to [JDF+11, WH08] for a comprehensive study of aesthetics and emotions in images.

Considering available datasets for evaluation, only a few small datasets exist today for affect / emotion analysis on visual content. A prominent one is the *International Affective Picture System* (IAPS) [LBC99] providing normative ratings of emotion (pleasure, arousal, dominance) for a set of color photographs. The dataset consists of 369 photos covering various scenes showing insects, puppies, children, poverty, diseases and portraits, which are rated by 60 participants using affective words. Similarly, the *Geneva Affective Picture Database* (GAPED) [DGS11] dataset provides 730 pictures including negative ones (spiders, snakes, scenes containing human rights violation), positive (human and animal babies, nature sceneries) and neutral pictures. All pictures were rated according to valence, arousal, and the congruence of the represented scenes. In Machajdik's work [MH10], the *Affective Image Classification Dataset* includes two separate datasets in the area of abstract painting (228 paintings) and artistic photos (807 photos) labeled by 8 basic emotions through a crowd-sourcing procedure.

Compared to the above works, the proposed approach in this chapter is novel and ambitious in two directions. First, it builds a large-scale ontology of semantic concepts reflecting a strong sentiment like "beautiful landscape" or "dark clouds" as a complement to a textual sentiment dictionary [ES06, WWH05]. Such an ontology is the first of its kind and opens new research opportunities for the multimedia and computer vision community. Additionally, in contrast to the above mentioned datasets the presented work provides a significantly larger dataset (about 500,000) of images crawled from social media and tagged with thousands of the ANP concepts. Furthermore, a separate image benchmark dataset from Twitter has been created in the context of this work, which aims exclusively for a sentiment prediction.

5.2.3 Visual Learning with Ontologies and Concept Combinations

As outlined in more detail in Chapter 2, the challenge of automatically detecting semantic concepts such as objects, locations, and activities in video streams - referred to as video annotation [ATY09], concept detection [SW09], semantic indexing [OAM+12] or multimedia event detection [CCC+11] - has been studied extensively over the last decade. In benchmarks like TRECVID [SOK06] or the PASCAL

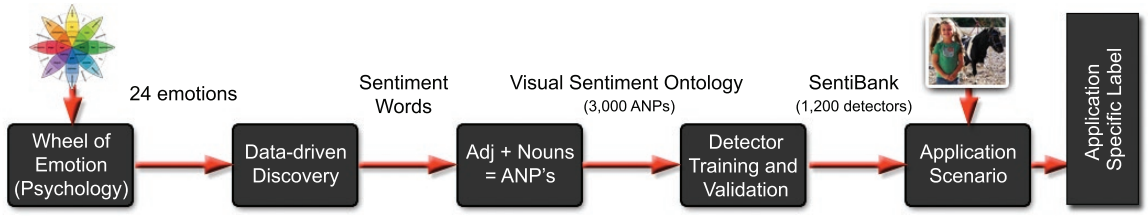


Figure 5.3: Overview of the proposed framework. As input a seed vocabulary in form of an emotional model from psychology is given. Once such a initial vocabulary is defined the construction of the Visual Sentiment Ontology (VSO) can be triggered. Founded on the final VSO a subset of concepts is trained and serves as detectors for SentiBank. Once such a mid-level representation is available, different application domains can be realized, which as input get an image or keyframe and compute a application scenario specific label as output.

visual object challenge [EVGW+10], the research community has investigated a variety of features and statistical models.

There has also been much work in creating large vocabularies and datasets such as ImageNet [DDS+09], consumer video [JYC+11], web video [TAP+10], or multimedia events [SMF+12]. Typically, such vocabularies are defined according to their utility for retrieval, coverage and diversity, availability of training material, and detectability by concept detection systems [NST+06, OAM+12]. Besides introducing large concept ontologies recent approaches have also turned towards new types of semantic concepts structures such as *bi-concepts* [LSWS12] or TRECVID's *concept pairs* [OAM+12]. These combined concept structures extend the idea of single-noun-concept to noun-pair-concept learning to enable search for two concepts co-occurring in visual content. For example, instead of constructing two single concept detectors such as a "horse" and "girl" detector to spot a "girl riding a horse", the idea is to train one "girl riding a horse" detector by harvesting social images being retrieved with "girl + horse" for detector training [LSWS12].

Also, over the last years a variety of practices have been proposed to bridge the semantic gap by the introduction of mid-level feature or attribute representations [FZ07, LNH09, FEHF09, KBBN09]. Such representations typically take the output of low-level features classifications as input for an additional learning of more specific target concepts. Examples include the learning of visual attributes [FZ07, BBS10, LNH09, FEHF09, YJT+12, RFF12], the construction of signatures from large concept detection vocabularies [HvdSS13, MHS13, TSF10] or the compilation of classifier banks such as ObjectBank [LSFFX10], DetectionBank [ASD12], or ConceptBank [MGvdSS13b].

The presented approach aligns with this thesis in the sense that a mid-level representation of visual content is introduced. The focus of this work, however, is less on supervised machine learning but rather on the construction of an ontology of visually detectable ANPs serving as mid-level representation of sentiment attributes in visual content. Compared to the above mentioned dataset collections or combined concept structures, which focus on generic concepts including objects, scenes, location (nouns only), the presented approach proposes novel adjective noun combinations allowing to capture sentiment visually. Although single concepts such as "magnetic drive" as found in the dataset definition of [SMF+12] could be considered as an adjective noun pair, the presented concept combinations in this chapter are motivated entirely differently by creating strong sentimental concepts and therefore adding notable value

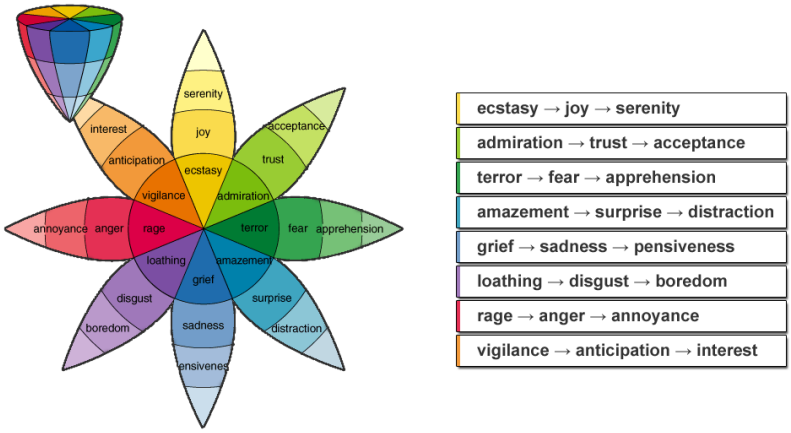


Figure 5.4: The psychological model used for ontology construction. Plutchnik's Wheel of Emotion [Plu80] with its 24 emotions organized by 8 basic emotions placed in a circle with each having 3 valences. These emotions serve as a seed vocabulary for the data-driven discovery of sentiment words.

as compared to "magnetic drive". Further it focuses on web-based tags for visual learning as employed in [KCK06, USKB10]. With respect to mid-level representations, this approach is to some extent related to the construction of detection banks, which are utilized for the detection of more complex concepts such as multimedia events. Also, to a certain degree, it shares similarity to ImageNet [RFF10] that is able to demonstrates the ability to learn attributes like "round", "furry" from its object ontology. Nevertheless the introduction of adjective noun combinations in this chapter is unique in terms of known mid-level representations.

5.3 Framework Overview

An overview of the proposed framework is shown in Figure 5.3. As input a seed vocabulary in form of an emotional model from psychological is given. Once such a initial vocabulary is defined the construction of the Visual Sentiment Ontology (VSO) can be triggered. Founded on the final VSO a subset of concepts is trained and serve as detectors for SentiBank. Once such a mid-level representation is available, different applications domains can be realized, which as input get an image or keyframe and calculate as output an application specific label.

5.3.1 Psychological Foundation

To establish a solid foundation for the construction of the VSO it is desirable to utilize a well-known emotional model derived from rigorous psychological studies. Besides early works such as Darwin's evolutionary motivation of emotions [Dar98], Ekman's facial expression system [E$^+$93] and Osgood's [OST57] appraisal and valence model, this work focuses on Plutchnik's cone like circumplex *Wheel of Emotions* model [Plu80]. As seen in Figure 5.4, the model is organized in 8 basic emotions, each having 3 valences being organized in a wheel like structure.

Argumentation for Plutchnik's Emotion Model

Plutchnik's model is inspired by chromatics and aligns emotions along a wheel placing bi-polar emotions opposite to each other. Since sentiment is characterized as bi-polar this property was found to be useful for the construction of a sentiment ontology.

Further, the model maps well to psychological theories such as Ekman, with 5 basic emotions are the same (anger, disgust, fear, sadness, surprise) and the last one, Ekman's "happiness" does not significantly differ from Plutchnik's "joy". Compared to the emotional model utilized in [MH10], Plutchnik basic emotions correspond to all 4 negative emotions and have slightly different positive emotions, which according to [MH10] does map well to Ekman's original work on facial expression. In contrast, Plutchnik introduced two additional basic emotions (interest, trust) and organizes each of them into 3 intensities providing a richer set of different emotional valences. Data statistics confirm the contributions of each emotion group in Plutchik to our final VSO as seen in Section 5.4.3.

5.3.2 Ontology Construction

As previously mentioned the ontology construction process is founded on psychological research such as *Plutchik's Wheel of Emotions* [Plu80]. During the first step - the data-driven discovery - for each of the 24 emotions defined in Plutchik's theory images and videos are retrieved from Flickr and YouTube respectively to extract concurrent tags (e.g., "joy" leads to "happy", "beautiful", and "flower"). These tags are then analyzed to assign sentiment values and to identify adjectives, verbs, and nouns. The set of all adjectives and all nouns is used to form adjective noun combinations or *Adjective Noun Pairs (ANP)* such as "beautiful flowers" or "sad eyes". Those ANPs are then ranked by their frequency on Flickr and sampled to form a broad and comprehensive ontology containing more than 3,000 ANP concepts. These ANP concepts serve as a foundation for detector training.

Adjective Noun Pairs

ANPs play a crucial role in the entire framework. Therefore some characteristics of adjectives, combinations of adjectives and adjective noun pairs are outlined next. The most essential purpose of their use in natural language is their particular capability to modify nouns. A noun changes its context entirely when combined with an adjective. Such modifying statements can be grouped into the following categories: opinion (beautiful, ugly, funny), size (large, small, tiny), age (young, old, ancient), shape (round, long, flat), color (yellow, blue, reddish), origin (German, western), material (wooden, metal), and purpose (barking dog). These groups also play an important role for the proper order of adjectives when – as common in the English language – used in combination with more than one adjective. For example one can say: "This is a beautiful sunny day". Such reinforced modifications of nouns, however, are not in the scope of the presented work.

Adjective Noun Pairs or *Paired Adjectives* as basic language elements have been known from linguistics since the learning about their association in the sixties [Pai63]. Recent work from Herman investigated semantic relationship between a noun and its adjectival modifiers in the context of its categorization, lexical context, and the efficacy of the latent semantic representation for disambiguation of word meaning [HBDP12]. Adjective noun pairs have also been used in the context of opinion analysis from textual online reviews as done by the *Review Spotlight* system [YNTT11].

5.3.3 Detector Bank Training

Based on the VSO and its representative dataset of Flickr images tagged with ANPs, a set of detectors is trained. For each of the 3,000 ANP, one single detector is trained and validated with respect to its performance. After selecting with reasonable performance a detector bank of as many as 1,200 ANP concept detectors is fixed. This *SentiBank* called library of detectors provides a 1,200 dimensional ANP detector response for a given image.

As illustrated in Chapter 2, a concept detection system consists of several feature classifier combinations. From this perspective SentiBank can be considered a concept detection system with 1,200 concept detectors, each comprised of a multi-feature classifier combination. The output of this detector bank can be interpreted as a multi-dimensional (i.e. 1,200 dimensional) encoding of ANP concepts presence for one image or keyframe. Following the detection output of one image or keyframe can be considered a vector describing concept probabilities for e.g. "crying baby", "beautiful sky", or "angry face".

5.3.4 Application Domains

The above described SentiBank output vector can be utilized in two ways: first the vector can be used as a visual explanation of content translating an image into a set of strongly responded ANPs. Second, it can be used as an input feature for an additional supervised classification based on a new dataset and different type of labels for training. For example to train a new detector for sentiment prediction one needs a dataset with sentiment labels to learn a mapping of the SentiBank output to in form of e.g. "crying baby" $\rightarrow$ "negative" and "beautiful sky" $\rightarrow$ "positive". This is also the motivation behind the understanding of SentiBank as a mid-level representation of visual content.

Given this fact, the application domains of SentiBank's ANP detectors are as broad as an appropriate labeled dataset is given. To demonstrate SentiBank's capability to generalize, the following application domains will be investigated in this chapter:

- **Sentiment Prediction:** SentiBank is used to predict sentiment values of image tweets to augment conventional sentiment prediction using text only
- **Emotion Detection:** SentiBank is used in the context of affective computing to detect different emotions in images.
- **Pornographic Filtering:** SentiBank is used to filter explicit adult content and illegal CSA material for law-enforcement support.

5.4 Visual Sentiment Ontology

In this section the design and systematic construction of the proposed Visual Sentiment Ontology (VSO) and its underlying image collection is outlined. In general, the sentiment being reflected by visual content is studied, i.e the perception by a human observer with respect to sentiment while looking at an image or a video. The construction process is founded on visual content shared on social media such as Twitter, Facebook, Flickr, or YouTube. The goal is to construct a large-scale ontology of semantic concepts, which *(1)* reflect a strong sentiment, *(2)* have a link to an emotion, *(3)* are frequently used and *(4)* have

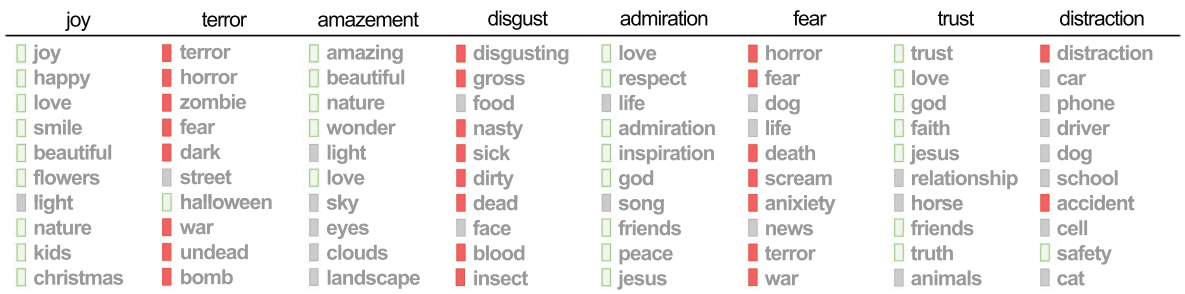


Figure 5.5: Examples for top tags for the emotions "joy", "terror", "amazement", and "disgust", "admiration", "fear", "trust", and "distraction". A green box indicates a positive sentiment, a grey box neutral sentiment and a red box a negative one. As seen some emotions lead to a retrieval of tags being exclusively positive or negative while some emotions lead to more neutral tags as output such as seen for "distraction".

reasonable detection accuracy. Moreover the VSO has the aim to be comprehensive and diverse enough to cover a broad range of different concept classes such as *people, animals, objects, natural or man-made places*, and so on. In the following each construction step will be explained in detail as illustrated in Figure 5.3.

5.4.1 Data-driven Sentiment Word Discovery

This section describes the extraction of sentiment words by automatically crawling Flickr and YouTube with the previously mentioned set of 24 emotions from Plutchnik's Wheel of Emotion as seed vocabulary. The goal of this procedure is to retrieve a large set of images and videos from these platforms for tag co-occurrence analysis derived from the given emotion queries.

Initial Image & Video Retrieval

For each of the 24 emotions a query is sent to Flickr and YouTube separately to retrieve images and videos. To retrieve images from Flickr their API is used with multiple search settings including title, description, and tags in combination with different time spans. In the case of video retrieval YouTube's API was used with separate category settings in combination with different time spans. An overview of the outcome from this retrieval procedure can be seen in Table 5.1 (a). The entire procedure was performed by the *Lookapp* tool [BUB11b] and led to a set of 310k retrieved media objects (150,034 images, 166,342 videos) in total. These images and videos were associated with 6.2M tags drawn from a set of 55k distinct tags. Although each tag might be associated with potentially multiple images or videos it can be seen that the set of distinct tags on Flickr is smaller than the one on YouTube (17,298 vs. 38,935) identifying the Flickr community as more consistent in the use of tags (not taking into consideration that video content might be more diverse and therefore requires a larger tag vocabulary). This way the initial seed vocabulary of 24 emotions can be seamlessly expanded by the tagging behavior of users on Flickr and YouTube, covering two major platforms for visual content sharing on the Internet.

Table 5.1: Statistics of the Visual Sentiment Ontology construction process. In **(a)** retrieval statistics from Flickr and YouTube are shown. In **(b)** sentiment word analysis statistics are listed, and in **(c)** statistics about the VSO and sample ANPs are given.

(a)	**Flickr**	**YouTube**	(b)	**Sentiment Words**
# of emotion queries	24	24	pos+neg adjectives	269
retrieved images or videos	150,034	166,342	neutral adjectives	0
tags	3,138,795	3,079,526	total adjectives	**268***
distinct tags	17,298	38,935	pos+neg nouns	576
avg. tags per image or video	20.92	18.51	neutral nouns	611
distinct top 100 tags	**1,146**	**1,047**	total nouns	**1,187***
distinct tags (both)	**1,771**		total verbs	**138***

*adjectives, nouns, verbs do not sum up to total distinct tags due to unknown word such as "xbox", "minecraft", or "vlog"

(c)	**VSO Statistics**
ANP concept candidates	320k
ANPs (non-empty image sets)	47k
ANPs included in VSO	**3k**
Strong positive sample ANPs	beautiful sky, little baby, happy family, sweet chocolate, nice beach
Strong negative sample ANP	dead animals, abandoned asylum, heavy storm, bad accident, scary bug
top positive adjectives	beautiful, amazing, cute
top negative adjectives	sad, angry, dark
top nouns	face, eyes, sky

Top Tags Analysis

Once the set of images and videos is retrieved for each emotion, an analysis of tag co-occurrences can be performed to automatically discover associations between emotions and user-generated tags. Prior to analysis stop-words are removed and stemming is performed on the raw tag meta-data. Next, for each set retrieved by an emotion query, an analysis is done and the top 100 tags for the set are ranked by their tag frequencies. Additionally, the sentiment value of each tag is computed using two popular linguistics based sentiment models, SentiWordNet [ES06] and SentiStrength [TBP+10]. In this chapter sentiment is computed as

$$s(w) \in \{-1, \ldots, 0, \ldots, +1\} \tag{5.1}$$

with $s(tag)$, the sentiment value of the word w having a continuous value from negative (-1) over neutral (0) to positive $(+1)$. Examples of such labeled top tags can be seen in Figure 5.5 for a subset of the 24 emotions. Interestingly, positive emotions tend to lead to positive sentiment words and vice versa. Nevertheless, there are also exceptions of emotions being associated with neutral tags as seen for the emotion "distraction" leading to neutral tags such as "car", "phone", and "dog" among others.

As a result, for all 24 emotions in total a set of 2,400 top tags was collected from both, Flickr and YouTube. It can obviously be the case that some tags might be retrieved by multiple emotion queries (e.g. "beautiful" might occur in the meta-data of images and videos retrieved by the query "joy" and

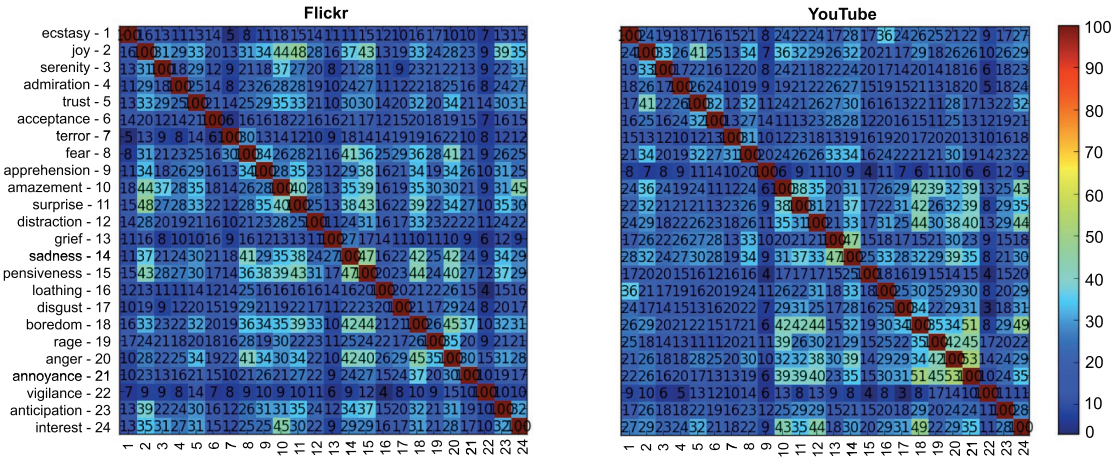


Figure 5.6: Co-occurrence matrix of tags being retrieved by multiple emotional queries based on the top 100 tags per emotion. There is a significant intersection of tags being used for emotion along the valence line of the emotional model use. Further some emotions tend to share tags as assigned by the users of the Flickr and YouTube platforms.

"serenity") the total number of distinct tags is much lower. The co-occurrence of the top 100 tags over all emotion is shown in Figure 5.6 color-coding the high overlap of shared tags between different emotions. It can be seen that while people do select different tags for the 8 basic emotions they prefer to choose similar tags for emotions along the valence lines (e.g "annoyance" and "anger" or "grief" and "sadness"). Also, some emotions such as "surprise" and "joy" or "fear" and "sadness" are linked through their intersection of underlying tags retrieved. As a final step each tag is categorized using Wordnet sysets, being either an adjective, verb or noun allowing to separate the set of tags into grammatical building blocks for the next step of the ontology construction. Overall, as shown in Table 5.1 (b), this procedure retrieved 1,146 distinct tags from Flickr and 1,047 distinct tags from YouTube forming the resulting set of 1,771 distinct tags for both with 1,187 nouns (576 positive and negative ones and 611 neutral ones) and 268 positive or negative adjectives. In general it can be said that during the retrieval and analysis more positive tags were found than negative ones.

5.4.2 Adjective Noun Pair (ANP) Construction

Looking at the results of the previous step it can be seen that the 576 nouns discovered with positive or negative sentiment would satisfy our initial condition *(1)* i.e. providing an ontology construction with strong sentiment but the remaining 611 neutral nouns would not, forcing them to be dismissed. Considering the adjectives, all 268 have either a positive or negative sentiment value (satisfying condition *(1)*) but probably they would not satisfy condition *(4)*: leading to a reasonable detection accuracy since visual learning of adjectives is difficult due to their abstract nature and high intra-class variability. To solve this dilemma, adjective noun combinations or *Adjective Noun Pairs (ANP)* are proposed to be the explicit semantic concept structure of the VSO. The advantage of ANPs, when compared to nouns or adjectives only, is the capability to turn a neutral noun into a strong sentiment ANP. Such combined phrases also make the concepts more detectable, as compared to adjectives only.

Table 5.2: Top ANPs for individual emotions from Plutchik's model given the described mapping procedure.

Emotion	**Top ANPs**
ecstasy	illegal drugs, hardcore techno, sexy legs
joy	happy smile, innocent smile, happy christmas
serenity	calm serenity, peaceful serenity, serene lake
admiration	fascinating places, charming places, excellent museum
trust	christian faith, rich history, nutritious food
acceptance	smooth curves, fat body, fat belly
terror	undead zombie, bloody zombie, creepy horror
fear	dangerous road, scary spider, scary ghost
apprehension	derelict farm, candid teen, scenic reserve
amazement	amazing talent, amazing scenery, amazing race
surprise	pleasant surprise, nice surprise, precious gift
distraction	drunk driver, noisy bird, excited child
grief	grieving mothers, funerary monument, funerary statue
sadness	sad goodbye, sad scene, sad eyes
pensiveness	lovely garden, misty road, young adult
loathing	illegal drugs
disgust	nasty bugs, dirty feet, ugly bug
boredom	tired feet, stupid sign, weird plant
rage	damaged road, noisy bird, drunk driver
anger	angry bull, angry chicken, angry eyes
annoyance	loud noise, sour apple, freezing morning
vigilance	-
anticipation	magical garden, tame bird, curious bird
interest	favorite architecture, great hall, fantastic architecture

Candidate Construction

The set of all positive and negative adjectives and the set of all nouns are now used to form ANPs such as "beautiful flower" or "disgusting food". After ANP concepts are formed, an extra text analysis on DBPedia [ABK+07] is employed to avoid ANPs, being named entities with changed semantics (e.g., "hot" + "dog" leads to a named entity instead of a generic concept or "dark" + "funeral" leading to a reasonable ANP but being associated with a dark metal band).

Obviously, during the construction of ANPs also the sentiment values of the adjective and the noun have to be fused. This is done by the idea of sentiment value reinforcement. Namely, to combine the corresponding sentiment values as following:

$$s(ANP) = s(adj) + s(noun), \quad s(ANP) \in \{-2, \ldots, +2\} \tag{5.2}$$

where $s(x)$ denotes the sentiment value of x. With this model, neutral nouns are colored by the sentiment values of the adjectives and strong sentiment values of adjectives and nouns are boosted such as in "cute" and "baby" both being positive to a very positive ANP ($s(cute\ baby) = +2$) or "bloody" and

"zombie" both being negative to a very negative ANP ($s(bloody\ zombie) = -2$).

However, in cases where adjective sentiment differs from noun sentiment the fusion of the combined ANP sentiment must be done carefully. For example, in cases like "abused" being negative and "child" being positive a straightforward fusion of sentiment values would obviously be wrong since both form the ANP "abused child" reflecting definitely a strong negative sentiment and not a neutral one. To address this issue, the presented system identifies cases whenever the sentiment of an adjective and a noun are of opposite value. The ANP then inherits the sentiment value of the adjective. As observed, in such cases the adjective usually has a stronger impact on the overall ANP sentiment than the noun.

Candidate Ranking

The outcome of the previous step leads to a set of 320k ANP candidates. These have to be filtered to remove meaningless or extremely rare constructions like e.g. "frightened hat" or "happy happiness". One of the goals of the VSO is to represent popular ANPs in social media. To reach this goal the filtering is done by frequency of ANP images found on Flickr and a subsequent ranking of all ANP candidates. Having this ranked list of ANP frequencies, which is characterized by a long tail as seen in Figure 5.7 (left)), all ANPs with no images found on Flickr are dismissed. The remaining 47k ANP candidates are taken as input for the final step, the sampling of the VSO. During this step cases are also eliminated where a singular and plural ANP both appear to become part of the VSO. In such a case the the more frequent ANP is taken and the other one is dismissed. For example, both nouns "dog" and "dogs" are retrieved and may form the ANPs "cute dog" or "cute dogs". However, because the ANP "cute dog" is more frequently used on Flickr the plural version of the ANP is dismissed and therefore will not become part of the VSO.

Ontology Sampling

The final sampling of ANPs is done under the condition to select only the most frequent and high sentiment value ANPs from the list of ANP candidates to form a broad and comprehensive ontology. For this, the frequency and sentiment values are fused together according to

$$score(ANP) = freq(ANP) * |s(ANP)| \tag{5.3}$$

where $freq(ANP)$ denotes the number of images found on Flickr. As a result the most frequent and strong sentiment ANPs are sampled from the overall set of 47k candidates. Focusing on adjectives, all candidate concepts are partitioned into adjective sets and from each adjective set its top n noun combinations of ANPs are taken. Further only ANPs with sufficient (currently set to > 125) images found on Flickr are considered. This guarantees the final VSO to be well balanced and diverse with respect to adjectives and only contain ANPs with at least a minimum popularity on Flickr.

The final VSO contains more than 3,000 ANP concepts being organized in 268 adjectives and their corresponding ANPs. Some of their strong positive APNs are: "beautiful sky", "little baby", and "happy family". The ANPs "dead animals", "abandoned asylum", and "heavy storm" are the negative counterparts. Adjectives with the highest number of noun combinations are "beautiful", "amazing" and "cute" representing positive adjectives and "sad", "angry", and "dark" representing the negative adjectives. In the same sense, nouns which are combined with the most adjectives are "face", "eyes", and "sky".

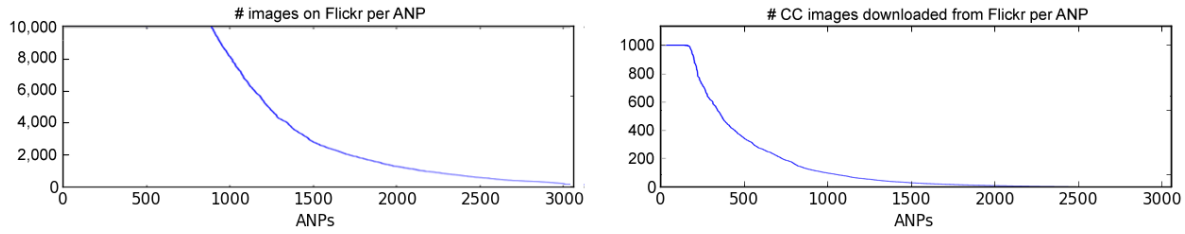


Figure 5.7: **Left:** Count of images on Flickr per ANP. **Right:** count of CC images downloaded per ANP. Please note that downloads were limited to max 1000 images per ANP.

5.4.3 Link back to Emotions

An interesting information is how the discovered ANPs are related to the emotions used in the very first retrieval step. To construct such a mapping first counts of images that have both, the emotion term and the ANPs string in their meta-data are retrieved. Theses values are then normalized to form the resulting 24 dimension histogram to sum one. This way a two-directional connection between an emotion and an ANP can be established. For example, the most dominant emotion for "happy smile" is "joy" and vice versa for the emotion "disgust" the ANP "nasty bugs". More examples can be seen in Table 5.2.

5.4.4 Flickr CC Dataset & Visualization Tool

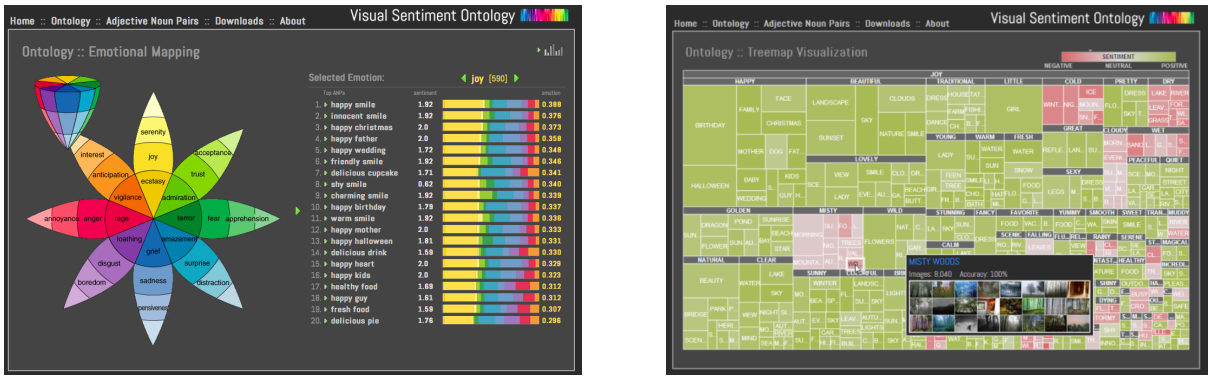
An essential part of the VSO is the retrieved dataset of Flickr images representing each ANP. The images are used as a dataset for SentiBank detector training (Section 5.6). Again the Flickr API was used to retrieve and download Creative Common (CC)[4] images for each ANP (limited to 1000 images) and ensure that only images are included that contain the ANP string either in the title, tag or description of the image. Following this condition sufficient amount of CC images were downloaded for 1,553 of the 3,000 ANPs (in total about 500k images). The distribution of the number of images can be seen in Figure 5.7 (right). Selected images of four sample ANPs are show in Figure 5.11 (left).

To help visualize the VSO and the associated large dataset, two novel visualization techniques were used, one based on the Wheel of Emotion (shown in Figure 5.8, left) and the other implementing the well-known TreeMap hierarchical visualization method (Figure 5.8, right). The Emotion Wheel interface allows users to view and interact with the Plutchik 24 emotions directly and then zoom in to explore specific ANP concepts and associated images. The TreeMap interface offers a complementary way of navigating through different levels of the VSO - emotion, adjective, noun, and ANPs. At each level, the map shows an intuitive visual summary of the number of images and the average sentiment value under each node. Interactive demos of these visualization tools of the proposed VSO are available online[5].

5.5 VSO Structure Construction and Analysis

So far the VSO consists of a list of ANP concepts derived by automatically mining Flickr and YouTube for tags being associated with one of the 24 emotions of Plutchnik's emotion model. However, an ontology

[4] `http://creativecommons.org`
[5] `http://visual-sentiment-ontology.appspot.com/`



(a) (b)

Figure 5.8: VSO visualization interface providing an emotion-to-ANP mapping (a) and a Treemap browser (b) which visualizes the entire ontology by navigating through different levels of the VSO including emotion, adjective, noun, and finally the ANP level.

does not only consist of concepts but also relations among them that can be used for browsing and reasoning about concepts within the ontology. To construct such ontological structures, an interactive process has been conducted in which multiple subjects were asked to combine concepts into distinct groups sharing coherent semantics among group members. The grouping process described is found on a separate consideration of adjective and nouns extracted from the list of ANPs. Separating adjectives and nouns allows the exploitation of relations unique for one of them which elsewhere would not have been observable. As a result of this construction process, a hierarchical structure of nouns (total of 520) was found to include 15 nodes at the top level with up to six levels depth. The adjectives (total of 260) were grouped to 6 nodes at the top level and two levels of depth. In these structures the standard hyponym-hypernym ("is-a") relations could be established in the noun hierarchy, while special relations like exclusive ("sad" vs. "happy") and strength order ("nice" vs. "great" vs. "awesome") were found among adjectives. Further a comparison of the constructed noun taxonomy and ImageNet [DDS+09] shows an overlap of VSO nouns and ImageNet synsets but also a significant amount of nouns being unique for the VSO, which are mainly related to strong emotions or sentiment as reported in more detail later in this section.

5.5.1 Methodology

To construct such ontological structures, an interactive process was conducted in which multiple subjects were asked to combine concepts into distinct groups. Each of such group should have a common semantic coherence and allow for a meaningful arrangement of an group members. Consensus among subjects was reached through result comparison and discussion. An overview of the construction process can be seen in Figure 5.9. First, the list of ANPs was split into independent sets of adjectives and nouns. From these two sets the resulting ontology was constructed, which consists of (i) an adjective grouping and (ii) a hierarchical taxonomy of nouns. The performed grouping for adjectives and nouns was done with the intention to allow exploitation of relations unique for adjectives or nouns.

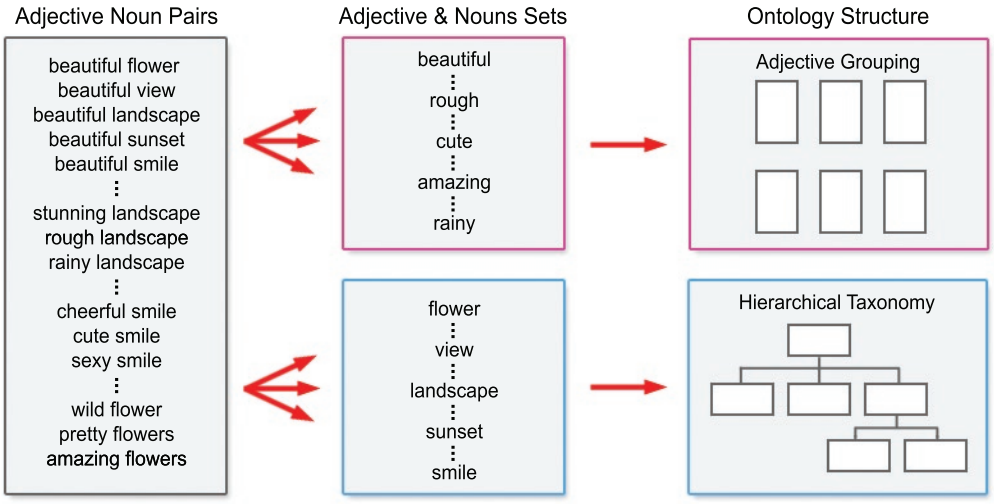


Figure 5.9: An overview of the ontology structure construction. First, the list of all ANPs is split into independent adjective and nouns sets. Second, starting from these sets for all adjectives a meaningful grouping is established and for all nouns a hierarchical taxonomy is created.

5.5.2 VSO Structure

The VSO ontology structure consists of two components, the adjective grouping and the hierarchical taxonomy of nouns. While a structure such as "dog" is a "pet" is an "animal" consists entirely of nouns from the VSO, a grouping for the following nouns "spring", "summer", "autumn", and "winter" would require the introduction of a *structure element* called `SEASONS`, which is not part of the VSO. Such exceptions, where new structure elements have to be introduced are indicated by all uppercase characters. In the following the two components will be described in more detail. A full overview of the distribution of adjectives and nouns and the number group members for the top level can be seen in Figure 5.10.

Adjectives

A total of 260 adjectives could be extracted from the list of ANPs. The subset of SentiBank adjectives was grouped to a two level structure with the following six nodes at the top-level: `WEATHER RELATED`, `OBJECT RELATED`, `LOCATION RELATED`, `PERSON RELATED`, `FOOD RELATED`, and `ANIMAL RELATED` adjective groups. Please note that although an adjective might belong to more than one group the best match for a grouping was chosen. For example, the adjective "lonely" might be used in combination with "lonely girl" or "lonely beach" either belonging to the group of `PERSON RELATED` or `LANDSCAPE RELATED` adjectives. However for this task "lonely" was chosen to belong to the group of `PERSON RELATED` adjectives. A complete overview of the groups can be found in the appendix in Figure B.1.

As seen in Figure 5.10 (left), the largest group of the VSO are `PERSON RELATED` adjectives, indicating a strong link to human related ANPs describing either the appearance (beautiful, ugly), characteristics (sick, drunk) or behavior (energetic, calm) of people. The second largest group of adjectives are `LOCATION RELATED` adjectives implying a place's emotional perception (safe, dangerous), characteristics (ancient, dry), or a individual impression (amazing, inspirational). It can be seen from this, that the VSO covers

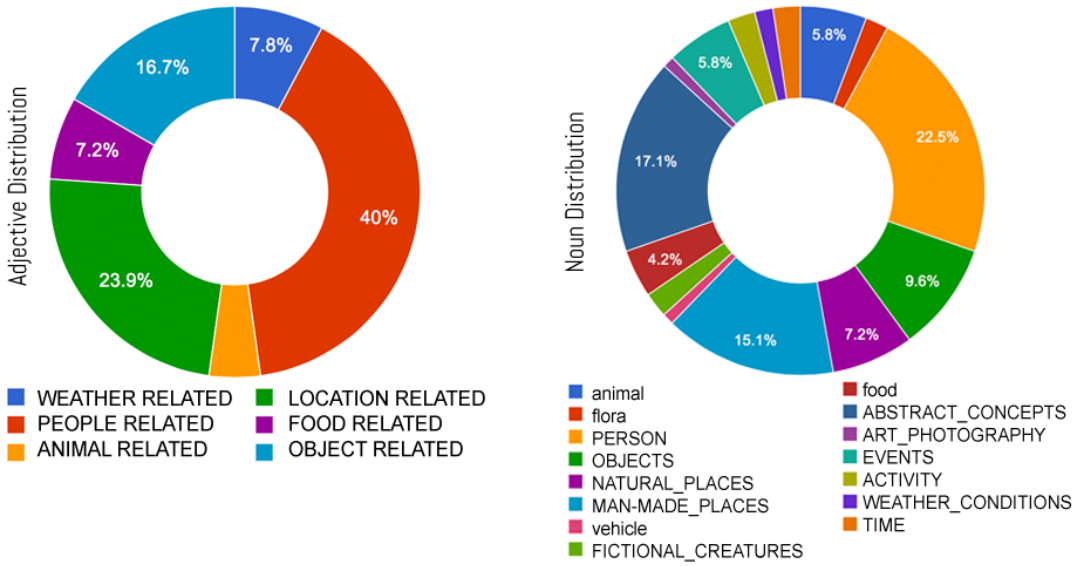


Figure 5.10: The distribution of sub-nodes for each of the top level nodes for **left:** adjectives and **right:** nouns of the VSO.

a broad range of different adjective groups on the one hand and is diverse enough within each adjective group to focus on different aspects of nouns. A detailed overview of the antonyms (e.g. "beautiful" $\leftrightarrow$ "ugly") found can be found in Section 5.5.3.

Nouns

The full hierarchical structure for nouns can be found in the appendix in Figure B.2. A total of about 520 nouns were found to construct a broader and deeper structure as compared to the adjective grouping. The hyponym-hypernym ("is-a") relations, which were discovered in the noun hierarchy have up to six levels of depth and 15 nodes at the top level. These top level nodes are either structural elements such as `PERSON`, `FICTIONAL CREATURES`, `MAN MADE PLACES`, `NATURAL PLACES`, `ACTIVITY`, `EVENTS`, `OBJECTS`, `WEATHER CONDITIONS`, `TIME`, `ART PHOTOGRAPHY`, and `ABSTRACT CONCEPTS` or nouns from the ANP list such as `vehicle`, `food`, `flora`, and `animal`.

As seen in Figure 5.10 (right), the derived hierarchy focuses on `PERSON` related nouns with sub-nodes such as specific groups (police, army, band, family), gender differentiation (girl, princess, widow, boy, king, actor), and body parts (belly, head, face) and face related actions (kiss, smile, tears) with groups, gender specific, body parts. However, the second largest top level node is `ABSTRACT CONCEPTS` being either related to an individual human (feelings, friendship, addiction) or to society (religion, security, freedom, art). This is followed by the third largest top level node: `MAN MADE PLACES`, with a broad range of outdoor (city, street, graveyard, chateau, house) and indoor (bathroom, hall, desk, bed) related nouns. From this observation it can be concluded that the VSO consists of both, nouns describing specific objects and nouns capturing comprehensive sceneries. A comparison to known ontologies such as ImageNet [DDS+09] and WordNet [Mil95] can be found in the next section.

Table 5.3: Adjective antonyms relations as found in the VSO. Such elements are of particular interest since they allow to form relationship of exclusion, which can be utilized for detector refinement.

Antonyms Relation
{sexy, attractive, beautiful, pretty, cute, adorable, handsome} ↮ {ugly, fat, chubby}
{evil, violent, angry, grumpy, rough, harsh} ↮ {innocent, friendly, helping}
{healthy, nutritious, fresh} ↮ {smelly, fat, greasy, rotten}
{laughing, smiling} ↮ {screaming, crying, grieving}
{sunny, clear} ↮ {cloudy, misty, rainy, stormy}
{colorful, bright, shiny, sparkling} ↮ {dark}
{delicious, yummy, tasty} ↮ {disgusting}
{peaceful, quite} ↮ {loud, noisy, barking}
{dry, dusty} ↮ {wet, slippery, muddy}
{energetic, excited} ↮ {tiered, sleepy}
{hot, warm} ↮ {cold, freezing, icy}
{healthy} ↮ {hurt, ill, sick, sore}
{terrible} ↮ {fantastic, excellent}
{sweet} ↮ {salty, bitter, sour}
{sad, lonely} ↮ {happy}
{scared} ↮ {calm, brave}
{empty} ↮ {crowded}
{safe} ↮ {dangerous}
{candid} ↮ {lying}
{curious} ↮ {shy}
{clean} ↮ {dirty}
{rich} ↮ {poor}
{young} ↮ {old}
{wild} ↮ {tame}

5.5.3 VSO Analysis

This section describes the characteristics of the previously introduced ontology structure with respect to adjectives and nouns. Once a structure is available its characteristics can offer valuable clues about the interplay and relations between ANPs.

Use of Structural Elements

As mentioned in the previous section, `STRUCTURAL ELEMENTS` have been introduced for the meaningful grouping of adjectives or nouns. This elements are not part of the ANP list of VSO but aim to organize the ontology. According to Figure B.1 and Figure B.2 from the appendix, there are 37 such elements as compared to a total of several hundred regular nodes in the ontology structure of the VSO. As already discussed, the hierarchical taxonomy of nouns is larger and more complex than the adjective grouping counterpart. Therefore, from those 37 elements, 6 belong to the top level nodes of the adjective groups and 11 to the top level nodes of the hierarchical taxonomy leaving the remaining 20 to be inner nodes of the noun hierarchy.

Table 5.4: Adjective Supportive Relation as found in the VSO. Such relations might be useful to reinforce detector scores from the SentiBank output.

Supportive Relation
delicious > yummy > tasty > healthy > nutritious > fresh > smelly > greasy > rotten
outstanding > incredible > amazing > stunning > awesome > great > nice
sexy > attractive > beautiful > pretty > cute > adorable > handsome
sunny > clear > cloudy > misty > wet > rainy > stormy
evil > violent > angry > grumpy > rough > harsh
hot > warm > cold > freezing > icy
dry > dusty > wet > slippery > muddy
friendly > pleasant > gentle > calm
creepy > haunted > scary > strange
hurt > bloody > ill > sick > sore
loud > noisy > peaceful > quiet
ancient > traditional > classic
gorgeous > charming > lovely
crazy > insane > mad
ugly > chubby > fat
fantastic > excellent
derelict > abandoned
laughing > smiling
damaged > broken
tired > sleepy
crowded > busy
stupid > dumb
little > tiny

Adjective Relations

As shown the set of adjectives from the VSO provides particular relations such as antonyms e.g. “cloudy” $\nleftrightarrow$ “sunny” and supportive properties such as “hot” > “warm”. An overview of such relations can be seen in Table 5.3 for antonyms and Table 5.4 for supportive relations. It can be seen that the antonym relations can be organized in sets of up to 7 elements, each adjective of the set being fully able to serve as an antonym for its adjective counterpart. Also, the number of positive and negative adjectives is imbalanced towards the positive ones. This observation goes hand in hand with the previously observed imbalance between positive and negative ANPs of the VSO. Regarding supporting relations or reinforcing relations, two conclusion can be drawn: First a link exists with the above mentioned antonym relations and second, they should be considered as a continuous strength signal useful for weighting rather than a binary exclusion one as implied for antonym relations. These relations play an important role in the assessment of ANPs and could be incorporated into the detection process to refine ANP detector responses. For example, the simultaneous detection of a “sunny sky” and a “cloudy sky” indicated the need for refinement in the final detection. Such a refinement can be understood as concept relation modeling as seen in Chapter 2.

Comparison to ImageNet and WordNet

Although the VSO has a specific purpose and utility, it is of interest to compare the VSO structure to known ontologies such as ImageNet and WordNet. WordNet [Mil95] is known as a large lexical database of the English language representing all adjectives, verbs, and nouns of the language as sets of cognitive synonyms (synsets). The purpose of WordNet is to support computational linguistics and natural language processing research. In contrast to this, ImageNet [DDS+09] aims to provide a large dataset of image samples for the subset of WordNet nouns.

Because the VSO puts an emphasis on visual content, the focus of the comparison is put on ImageNet treating WordNet as the superset of all elements of natural language. Please note that, while the VSO does consist of adjective and noun combinations, the comparison with ImageNet is based on the set of all nouns extracted from ANPs. The comparison of the constructed noun taxonomy and ImageNet with its 21,841 synsets shows that 59% of the VSO nouns can be mapped to ImageNet synsets by comparing synset names with noun nodes of the VSO structure. This leads to 41% of VSO nouns not being covered by ImageNet, although being found in WordNet. These concepts unique to VSO, are mostly related to abstract concepts such as "violence" or "religion", which reflect strong emotions or sentiments. This confirms the unique focus on emotions and sentiments in the concept discovery process of VSO, as described earlier in this chapter.

5.6 SentiBank

Derived from the Visual Sentiment Ontology constructed above, *SentiBank*, a novel sentiment classification framework, is proposed. It encodes the output of ANP detectors into a mid-level concept representation. SentiBank's objective is to detect ANP concept presence and to characterize the sentiment reflected in visual content (although in the next section it will be shown that SentiBank's capabilities are not limited to sentiment prediction only). In this section several key issues regarding the construction of SentiBank are addressed. First, ANP label reliability will be discussed, then the design of individual ANP detectors and their detection performance are reported. Finally the usage of special features within SentiBank will be outlined.

5.6.1 Reliability of ANP labels

It is well known that web labels (image or video) may not be a reliable indicator or concept presence [UBB10, DDS+09, USKB10] (Chapter 4 covers this issue). Since the underlying dataset for SentiBank's detector training is acquired from Flickr and ANP labels are given by Flickr users, a setup is given, where such *pseudo labels* might be either "false positives", i.e. an image is labeled by an ANP but actually does not have the ANP in the image content or "false negatives", i.e. if an image is not labeled by an ANP it does not automatically imply the ANP is not present in the image. For instance, an image labeled with "cloudy sky" may not show a cloudy sky and vice versa, an image labeled with "beautiful girl" may also show a "happy face", even though the label has not been given to the image.

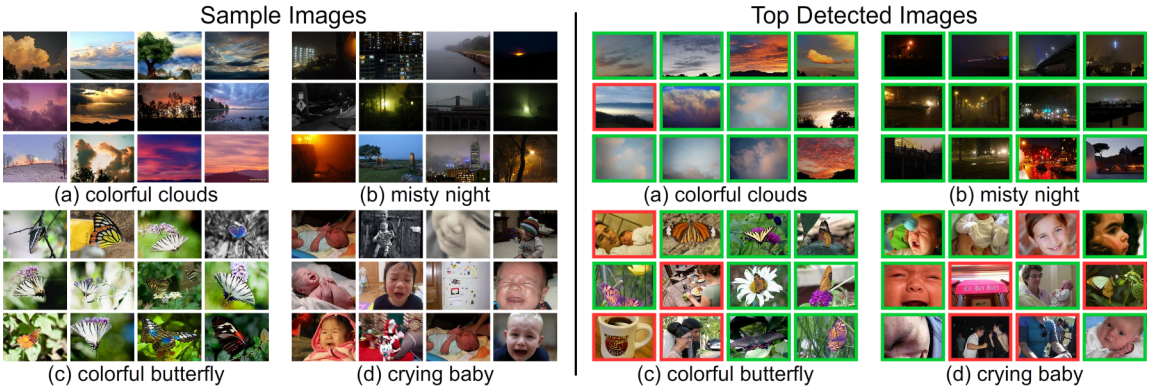


Figure 5.11: **Left:** Selected images for four samples ANPs, **(a)**,**(c)** reflecting a positive sentiment and **(b)**, **(d)**, a negative one. **Right:** top detected images by SentiBank ANPs with high detection accuracy (top) and low accuracy (bottom). Correct detections are surrounded by green frames and incorrect ones are surrounded by red frames.

Dealing with Pseudo Web Labels

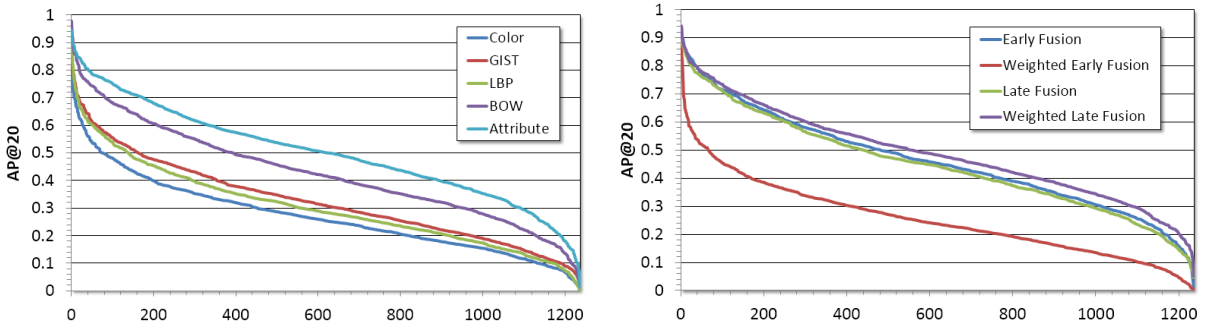
Considering pseudo positive labels and the potential of *false positives*, the reliability of ANP labels is evaluated by an Amazon Mechanical Turk (AMT) experiment[6]. Sample images are randomly sampled from 200 ANP concepts to manually validate their image labels, namely using AMT Turker judgment to check whether an image indeed contains the corresponding ANP. Each image label is validated by 3 Turkers and is treated as "correct" only if 2 Turkers agree that the image is showing the given ANP label. Results of this experiment show that a very high percentage (>90%) of AMP image labels are actually "correct", which indicates that the careful selection of images which indeed contain the ANP name in their title, description, or tags lead to a low fraction of false positive ANP labels.

Unfortunately, on the *false negative* case, such a label validation is too exhausting to be feasible since it would require to fully label all images for all ANPs asking to perform roughly 1.5 Billion label judgments[7] in total. This is also an open issue for existing crowdsourced visual recognition benchmarks such as ImageNet ILSVRC2010-2012[8], ObjectBank [LSFFX10] and Classemes [TSF10]. Recently, in ILSVRC2013, researchers have also started to fully label the presence / absence of all synsets in every test image. To deal with this issue, the negative class is randomly sampled from positives of other ANPs except the ones that are highly related, such as ANPs with the same adjective or noun. This way the probability to include a false negative can be minimized, while avoiding the prohibitive task of labeling the entire dataset for every ANP concept.

Training and Testing Partition

The training set for each ANP is sampled with 80% of positive pseudo labeled images (on average 256 pseudo positive images per ANP) of the ANP class and twice as many negative ANP samples using the subsampling scheme described above. For testing, two different test sets are prepared, denoted as the *full*

[6] `https://www.mturk.com/mturk/`
[7] need to verify 3,000 ANPs pseudo labels in each of the 500,000 images
[8] `http://www.image-net.org/challenges/LSVRC/2013/index`



(a) individual detectors (b) fusion schemes

Figure 5.12: **Left:** Comparison of ANP detectors using different features. Performance computed by averaging over 5 runs of the reduced testsets. **Right:** AP@20 of ANP detectors using different fusion approaches. Performance computed by averaging over 5 runs of the reduced testsets.

and *reduced* test sets (on average 64 pseudo positive images per ANP). Both use the remaining 20% of pseudo positive samples of a given ANP as positive test samples. But the amount of negative samples are different - the full testset includes 20% pseudo positive samples from each of the "other" ANPs (except those with the same adjective or noun). This leads to a balanced training set and a large and diverse test set for individual detector performance evaluation. However, the prior of the positives in each test set is very low, only about 1/1,553. The reduced testset, intended for fast implementations and balanced test sample distributions, includes much less negative samples - the number of negative test samples for each ANP is just twice as many as the positive samples. To avoid test set bias, also 5 runs are performed of the reduced test set, each of which includes different negative samples while keeping the positive samples fixed. This arrangement will be used for later experiments and performance will be averaged over these 5 runs (Figure 5.12 and Figure 5.13) in this chapter.

5.6.2 ANP Detector Training

Once the partition of the dataset into training and test sets is done, detector training can start. For each ANP from the VSO a detector is trained according to the following concept detection pipeline setup.

Visual Feature Design

Following the feature design for state-of-the-art visual classification systems such as ObjectBank [LSFFX10] and Classemes [TSF10], the following generic visual features for ANP detector training are employed: a 3×256 dimensional Color Histogram extracted from the RGB color channels, a 512 dimensional GIST descriptor [OT01] since a significant proportion of ANPs relate to scenes like "beautiful landscape", a 53 dimensional Local Binary Pattern (LBP) descriptor suitable for detecting textures and faces, a Bag-of-Words quantized descriptor using a $1,000$ visual word dictionary with a 2-layer spatial pyramid and max pooling, and finally a $2,659$ dimensional Classemes descriptor [TSF10] to characterize abstract ANPs. Additional features specialized in detecting objects, faces, or aesthetics will be presented later in Section 5.6.4.

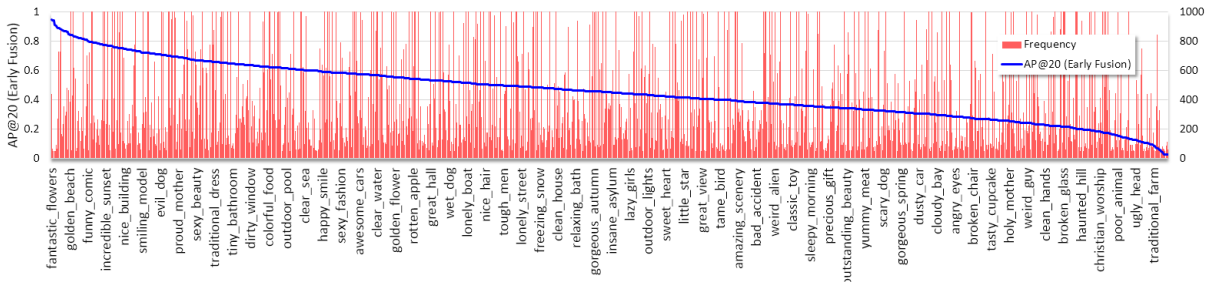


Figure 5.13: AP@20 (average over 5 runs of the reduced testsets) vs. frequency of 1,553 ANP detectors ranked by detector performance. It can be seen that the performance of the detector is not tightly bound to the frequency of available training samples. Note only a subset of ANP names are shown due to limited space.

ANP Detector Training and Evaluation

Due to the large amount of ANPs in the ontology, Linear SVMs are utilized to train ANP detectors and to ensure high efficiency. Parameter tuning of the SVM was performed by a 5-fold cross-validation optimizing Average Precision at rank 20 (AP@20), a performance measure focusing on the accuracy of the top ranked samples. Detector performance was measured also by the Area Under Curve (AUC), describing the probability to rank a random positive sample higher than a random negative one. Finally, the third measure is the F-Score, describing the harmonic mean between precision and recall. All three measures are considered standard measures for detector evaluation as seen in Chapter 2.

Results of detector performance using various features can be seen in Figure 5.12 (left). Here a clear dominance by the attribute features followed by Bag-of-Words (BOW) can be observed. Considering feature fusion, both early and late fusion schemes are evaluated. The former refers to merging and normalizing different feature vectors into a single vector. The latter refers to the fusion of detector scores after classification. Figure 5.12 (right) illustrates the results of different fusion methods including *Early Fusion, Weighted Early Fusion, Late Fusion, and Weighted Late Fusion*[9]. It can be seen that Weighted Late Fusion out-performs other fusion schemes by a small margin, while the performance of early and late fusion is quite close. For implementation simplicity, an early fusion approach will be used in the released SentiBank detectors.

5.6.3 SentiBank Construction

The described training step is performed for each individual ANP. After training each detector is tested against the given testset and its performance is evaluated. Based on this evaluation SentiBank is constructed.

ANP Detectability Overview:

An important step in building SentiBank is to select only ANPs with reasonable detection accuracy. First, the ANP detectors are ranked based on the previously described performance measures such as F-Score, AUC, or AP@20. It is worth to note that selections based on F-Score, AUC, or AP@20 only

[9] weights are also tuned by cross-validation

slightly affect the relative orders of ANPs. Then the top $1,200$ ANP detectors are selected, all of which have non-zero AP@20 and most have an F-score greater than 0.6, when evaluated over the reduced testset.

It is interesting to see (as shown in Figure 5.13) that there is no correlation between the detectability of an ANP and its occurrence frequency. Instead, the difficulty in detecting an ANP depends on the content diversity and the abstract level of the concept. Figure 5.11 shows some examples of the best and worst performing ANPs based on AP@20.

5.6.4 Special Visual Features

As reported in [BJC+13] also several special features have been tested for training the SentiBank detectors. First, since many of the ANP concepts are associated with objects, object detection techniques are utilized to localize the concept within an image. For the entire set of SentiBank detectors 210 ANPs are chosen that are associated with detectable objects such as people, dogs, or cars. Having these ANPs identified, the object detection tools from [LSFFX10] are applied and combined with multi-scale detection results to form a spatial map constraining the image region from which the visual features described in Section 5.6.2 are extracted. Another option is to take the object detection response scores directly as features. Doing so, facial features on 99 ANPs with nouns like face, smile, tears, etc. are evaluated. These include the detected face count, relative face size, relative facial marker position and Haar features at the markers. Thirdly, aesthetics related features on the entire list of ANPs are tested. These features from [BSS11] include dark channel and luminosity features, sharpness, symmetry, low depth of field, white balance, etc. The above three groups of features increase the mean AP@20 score of selected ANP sets by 10.5%, 13.5% and 9.0% (relative gains) respectively on the reduced testset and the mean AP@100 by 39.1%, 15.8% and 30.7% on the full testset. Based on the above comparisons, the conclusion can be drawn that generic features offer a competitive solution for detecting ANP visual sentiment concepts, while special features offer great potential for further improvements.

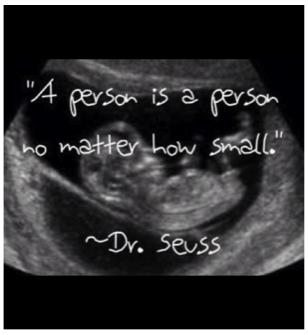
5.7 SentiBank Applications

In this section several applications of SentiBank are presented. Since the initial motivation to construct the VSO and create SentiBank is to capture the sentiment reflected in visual content, the first evaluation of SentiBank focuses on sentiment prediction in image tweets as an application domain.

Nevertheless, since the potential of SentiBank as a mid-level representation of visual content reaches beyond sentiment prediction two additional application domains are presented. They include emotion classification against a well-known emotion dataset of art photos [MH10] and the detection of pornography in general and filtering of CSA material in particular.

5.7.1 Sentiment Prediction

With respect to sentiment prediction, state-of-the-art approaches typically rely on text-based tools such as SentiWordNet [ES06] or SentiStrength [TBP+10]. However, due to the length restriction of 140 characters in tweets, such approaches are challenged by the short amount of text. Even humans are often unable to correctly discern the sentiment of the text content as seen in the beginning of this



(a) True stuff. I have mad respect for all the ladies that DO NOT give in to abortion.

(b) #groundzero #hurricanesandy #newjersey

(c) ouch mr police man

Figure 5.14: Three random sample tweets from the photo tweet dataset with their photos and textual content.

chapter by the challenging tweet examples from Figure 5.1. To overcome this issue, SentiBank is used to complement and augment the text features with visual analysis for sentiment prediction.

Photo Tweet Sentiment Benchmark Dataset

Unfortunately, no dataset exists in the literature which puts a particular emphasis on sentiment prediction from photo tweets i.e. textual tweets containing a shortened URL pointing to a photo. Examples of such tweets can be seen in the initial example of this chapter (Figure 5.1). In the context of sentiment prediction such a dataset should provide not only photo tweets but also annotations with labels such as "positive", "neutral", or "negative" or related sentiment valence scores. Because of the non-existence of such a benchmark dataset, the first step in the evaluation of SentiBank is the creation of a photo tweet dataset including data acquisition and ground truth labeling. Please note that although sentiment is often specific to a particular domain such as movies, food, or politics, for this study a generic approach covering the broad spectrum of tweet topics is adopted.

Hashtag Selection The benchmark dataset created for the experiments in this section is retrieved using the PeopleBrowser API[10] by collecting tweets containing photos according to the following popular hashtags:

- **Human:** #abortion, #religion, #cancer, #aids, #memoriesiwontforget
- **Social:** #police, #nuclearpower, #globalwarming, #gaymarriage,
- **Event:** #election, #hurricanesandy, #occupywallstreet, #agt (america got talent), #nfl, #blackfriday, #championsleague, #decemberwish
- **Person:** #obama, #zimmerman
- **Location:** #cairo, #newyork,
- **Technology:** #android, #iphonefan, #kodak, #androidgame, #applefan.

[10] `https://www.peoplebrowsr.com`

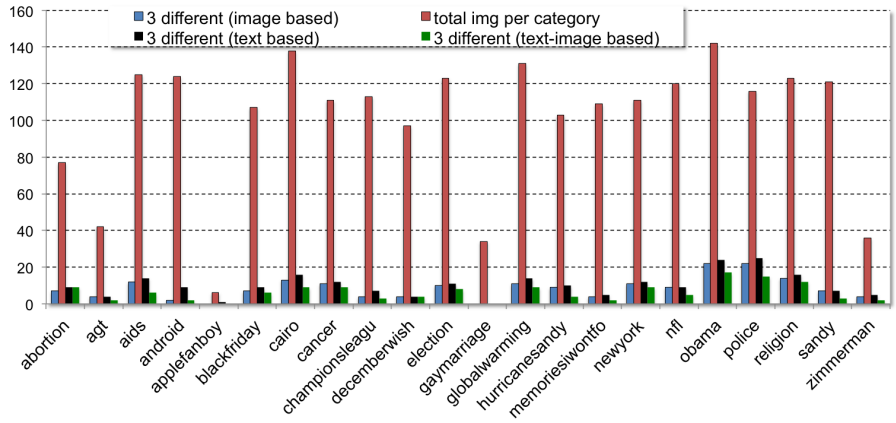


Figure 5.15: The volumes and label disagreements for different hashtags. For each hashtag, the total number of images is shown, in addition to the number of images receiving complete disagreement among turkers (i.e., 3 different sentiment labels: positive, negative and neutral), while labeling is done using text only, image only, and joint image-text combined.

The spectrum of hashtags for retrieval of photo tweets covers a broad range of topics starting from human, social, event, person, location, to technology related areas. As seen, the set of hashtags selected for retrieval aims at controversy topics which have a higher probability of providing polarizing opinions. The resulting dataset consists of 20 to 150 images per hashtag (in total 2,115 images were collected), which were crawled during August 2012. Some random example photo tweets can be seen in Figure 5.14. As illustrated these photo tweets are characterized by significantly less text than the regular 140 character restriction given by Twitter. They also only list different hashtags or do not provide textual descriptions at all.

Ground Truth Labeling: Up to now an open question is if, considering ground truth labeling, photos have the potential to support sentiment prediction if textual data is already available i.e. does the inclusion of photos provide additional help for humans in judging the sentiment of a photo tweet. To obtain sentiment ground truth for the collected image tweets, three labeling runs have been conducted using AMT, namely *image-based*, *text-based*, and joint *text-image based* runs. They correspond to image-only inspection, text-only inspection, and full inspection of both image and text contained in the tweet. For each labeling run, 3 randomly assigned Turkers are asked to label text, an image or the entire image tweet independently i.e. no Turkers are asked to annotate the same tweet under different modality settings. Figure 5.15 shows the labeling statistics, where an image is defined as "agreed", if more than 2 Turkers assign the same label (either positive, negative or neutral). From the result it can be clearly seen that, joint *text-image based* labels are the most consistent ones, followed by *image-based* labels and then the *text-based* labels. This indicates the limitation of text-based sentiment analysis for Twitter if photos are involved and highlights the potential for a holistic sentiment analysis using both the image and text analysis. However, in the end, only the photo tweets are included in the benchmark that receive unanimously agreed labels from three Turkers of the joint image-text annotation as the final benchmark set. It includes 470 positive tweets and 133 negative tweets over 21 hashtags, among which 19 hashtags each with more than 10 samples are shown in Figure 5.16.

Table 5.5: Comparison of Tweet Sentiment Prediction Accuracy. Results are illustrated in a matrix for comparison of textual, visual, and combined sentiment prediction accuracy. Evaluated systems are shown at the left ranging from (1) to (8) with SentiBank outperforming all evaluated baselines.

		text only	**visual only**	**SentiStrength + SentiBank**
(1)	Naive Bayesian	0.57	-	-
(2)	SentiStrength	**0.61**	-	-
(3)	Low-level features + Linear SVM	-	0.55	-
(4)	Low-level features + Logistic Regr.	-	0.57	-
(5)	SentiBank + Linear SVM	-	0.67	-
(6)	SentiBank + Logistic Regr.	-	**0.70**	-
(7)	SentiBank + Linear SVM	-	-	0.68
(8)	SentiBank + Logistic Regr.	-	-	**0.74**

Text Based Classification Baselines:

First, text only sentiment prediction baselines are established. These are used as comparison to evaluate image sentiment prediction performance and joint text and visual sentiment prediction performance. Two text-based sentiment predictor baselines are adopted: `(1) Naive Bayesian text-based Sentiment Classifier:`

$$Score = \frac{1}{M} \sum_{m=1}^{M} Frequency_m \times Score_m \tag{5.4}$$

in which *Score* is the sentiment prediction score normalized to [-1,1], M the number of unique words after stemming and stop word are removed, $Frequency_m$ the frequency of word m, and $Score_m$ is the individual sentiment score of word m obtained from SentiStrength. `(2) SentiStrengh API:` To directly leverage state-of-the-art sentiment prediction the publicly available SentiStrength API[11] is used as baseline. Here, for the entire tweet text a sentiment score is retrieved.

Baseline results can be seen in Table 5.5 (lines 1-2). It can be seen that SentiStrenght API perdition accuracy based on the entire tweet text is higher than the one combining scores of individual words using the Naive Bayesian method.

Visual-based Classification Performance:

As mentioned before, SentiBank serves as an expressive mid-level representation of visual concepts. For each image SentiBank provides a 1,200 dimensional ANP response, which is used as an input feature for the sentiment classification. Here, classifiers such as Linear SVM and Logistic Regression are employed. To this end, the aim is not only to predict the sentiment being reflected in images but also to provide an explanation of the given prediction. This is achieved by providing a list of top responding ANP detectors in addition to the sentiment prediction label.

First, the proposed SentiBank mid-level representation is compared with low-level features (the same, which were used for ANP detector training) using two different classification models, LinearSVM and Logistic Regression. For low-level features, the same setup is used as those described in Section 5.6.2

[11] `http://sentistrength.wlv.ac.uk/`

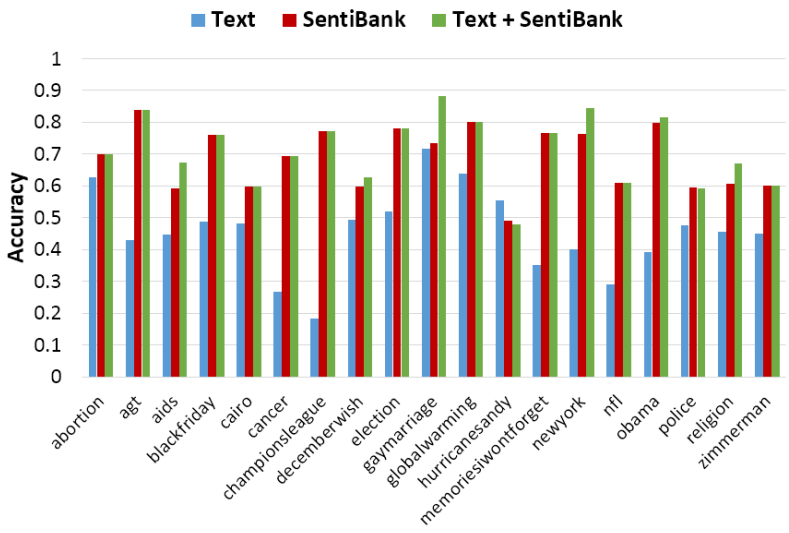


Figure 5.16: Photo tweet sentiment prediction accuracy over different hashtags by using text only (SentiStrength), visual only (SentiBank), and a combination of both. Accuracy is averaged over 5 runs.

(color histogram, GIST, LBP, BoW, and attributes). Prediction accuracy is shown in Table 5.5 (lines 3-6) and confirms the significant performance improvement (more than 20% relatively) achieved by the SentiBank features. The logistic regression model is also found to be better than Linear SVM.

Joint text-image Classification Performance:

Finally, a combined sentiment prediction accuracy is presented and compared to the previous setups. Here, SentiStrength is used in a late fusion setup with SentiBank to predict sentiment labels. As previously seen visual based methods using SentiBank concepts are significantly better than the text only ones (70% from line 6 vs. 61% from line 2 in Table 5.5). By further analyzing the results, one can recognize that most of the text contents in the tweets are short and neutral, explaining the low accuracy of text-based methods in predicting the sentiment. In such cases, the sentiment values of the visual content predicted by the SentiBank-based classifiers play a much more important role in predicting the overall sentiment of the tweet. However, when further combining both systems to a joint image-text based sentiment prediction a significantly better performance can be achieved than visual-only or text-only, by 4% and 13% absolute gains respectively (Table 5.5, lines 7-8).

Figure 5.16 shows a comparison of sentiment prediction accuracy for each individual hashtag. Here, its can be seen that the visual-based approach using SentiBank concepts consistently outperforms the text-based method using SentiStrength API, except for one hashtag ("hurricanesandy"). It is also very encouraging to see combining text and SentiBank features further improves accuracy for several hashtags, despite the low accuracy of the text-based method.

Results of a few sample images can be seen in Figure 5.17. Here, SentiBank's capability in explaining predictions is illustrated by showing a list of top ANPs detected in each test image.

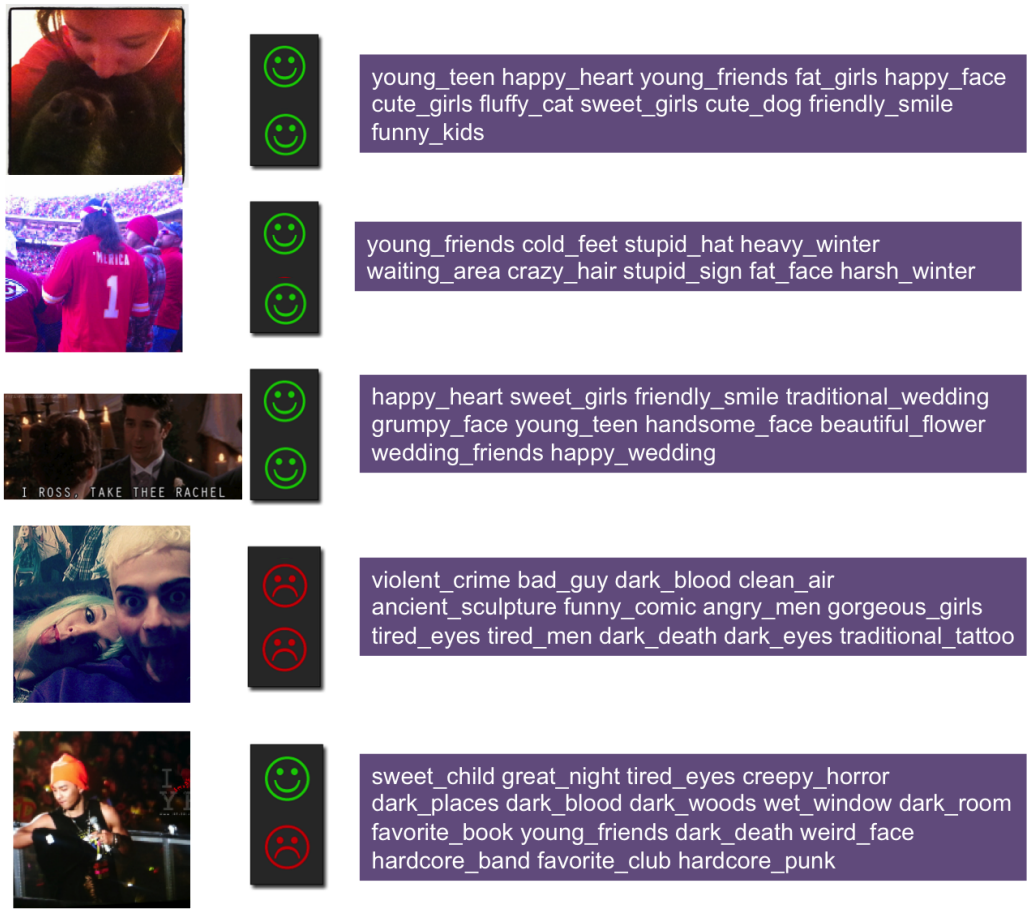


Figure 5.17: Sentiment prediction results for sample image tweets using SentiBank as features (top icon: ground truth sentiment, bottom icon: predicted sentiment). On the right the top responding ANPs found in the sentiment prediction model.

5.7.2 Emotion Classification

Although the initial motivation for SentiBank was to predict sentiment reflected in images, a comparison of the proposed method performing emotion classification might be of interest. Here, SentiBank is evaluated for emotion classification and compared to the performance of [MH10]. The dataset in [MH10] is based on ArtPhotos retrieved from DeviantArt.com and contains 807 images covering 8 emotion categories. This kind of evaluation poses a set of challenges such as the domain change i.e. SentiBank is trained on a different set of images than the test set. Moreover, the emotion categories are slightly different, not mentioning the emphasis of SentiBank as a framework with generic visual features rather than the specialized affective features used in [MH10]. During the evaluation a similar process to select features is followed for each emotion category by using the weights of individual features, combined with the Naive Bayesian classifier. Results are reported in Figure 5.18. Even in such a challenging setting, SentiBank compares relatively well to [MH10] when using the same classification model (Naive Bayesian) and even slightly outperforms the best results in [MH10] when using the presented Logistic Regression model. This illustrates the applicability and potential of SentiBank for applications in different domains.

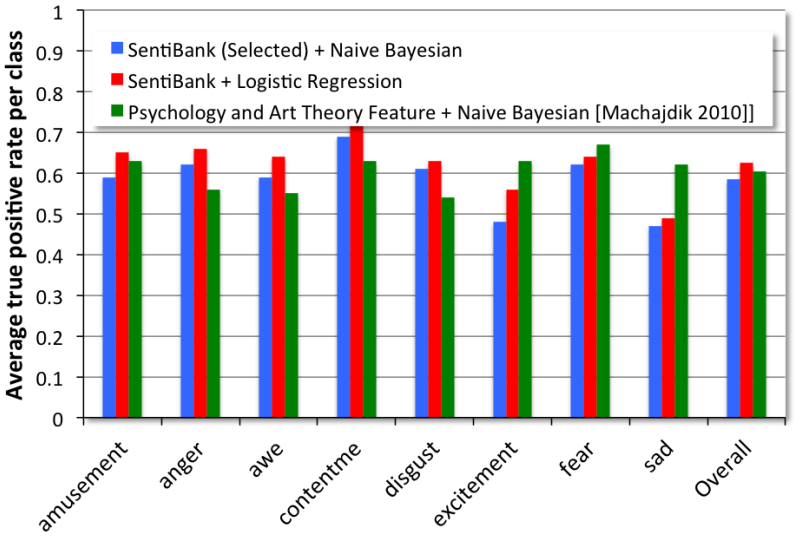


Figure 5.18: Emotion classification performance compared with [MH10] using the ArtPhoto dataset.

5.7.3 Digital Forensics

Another application scenario for SentiBank is the filtering of illegal pornographic content to support law enforcement during investigations. In particular, this includes the content-based detection of regular pornography [JUB09, USBS12] and the detection of child sexual abuse (CSA) material [US11]. In this context SentiBank is tested to detect *Adult vs World*, *CSA vs World*, and *CSA vs Adult* content. In collaboration with police partners and European cyber-crime units, experiments on three datasets were conducted. Each datasets consists of 20,000 images with "World" images being randomly downloaded from Flickr; "Adult" images being acquired from explicit pornographic websites including different categories such as *amateur*, *mature*, *teen*, and others; and finally "CSA" images being provided by law enforcement as real world illegal child pornographic classified content[12].

SentiBank was evaluated against several low-level features such as color-correlograms [HKM+97], visual words [SZ03], visual pyramids [LSP06] and a skin detection [DPN08], commonly used in pornography detection. For classification SVM with RBF and χ^2 kernel were used (parameters were optimized by 5-fold cross-validation).

Detection results can be seen in Table 5.6. The evaluation reveals the potential of SentiBank for this application scenario. As seen, the utilization of SentiBank features shows the best performance for the *Adult vs World* and *CSA vs World* test runs. For the very challenging *CSA vs Adult* setup, SentiBank perform similar to the best performing system using colorcorrelogram features. In all three test runs a late fusion of features could further improve detection with SentiBank, always providing the largest contrition in terms of fusion weights. However, when compared to low-level features SentiBank has the advantage of explainability, a property often requested by low-enforcement. Due to the ANPs assigned to each SentiBank score, the results of single image classifications lead to the following insights about the characteristics of pornography and CSA material as seen in Figure 5.19. The results show that for

[12]Given the illegal nature of CSA data only numeric feature information was provided by law enforcement.

Table 5.6: SentiBank feature performance and fusion results adding important low-level features. The learned weights indicate that SentiBank provides the most valuable information, except for CSA vs. Adult where it contributes equally with color-corellograms.

	Adult vs World		CSA vs World		CSA vs Adult	
Feature	**AVP**	**EER**	**AVP**	**EER**	**AVP**	**EER**
SentiBank	**0.9715**	**0.0904**	**0.9712**	**0.0832**	0.8996	0.1746
colorcorrelogram	0.9531	0.1075	0.9293	0.1403	**0.9107**	**0.1683**
vispyramids	0.9510	0.1145	0.9136	0.1608	0.8741	0.2058
viswords (dense)	0.9453	0.1208	0.9138	0.1613	0.8758	0.2028
skin segment	0.9242	0.1365	0.8132	0.2643	0.7424	0.3360
fused	**0.9797**	**0.0726**	**0.9781**	**0.0622**	**0.9470**	**0.1080**

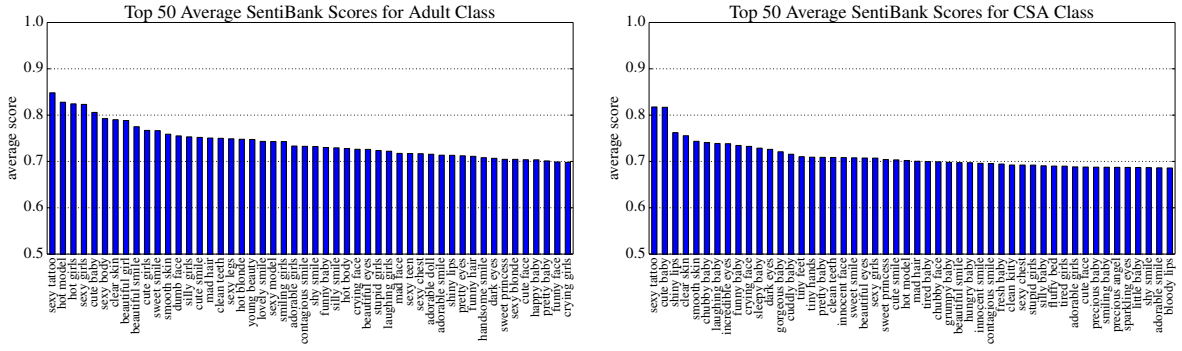


Figure 5.19: Top 50 responding SentiBank detectors, averaged over 20,000 pornographic (l) and child pornographic (r) images. The ANPs for both classes reveal a prominent relation to body parts (eyes, chest, teeth, legs, feet). Additionally, top ANPs for adult class images often contain the term *girl*, while for CSA class images *baby* appears more frequent indicating a child's presence.

adult material the adjectives "hot" and "sexy" are often present, while "cute" and "tiny" appear frequently for CSA content. The ANPs for both classes reveal also prominent relation to body parts (eyes, chest, teeth, legs, feet). Additionally, top ANPs for adult class images often contain the term *girl*, while for CSA class images *baby* appears more frequently indicating a child's presence.

5.8 Discussion

In this chapter an approach towards the prediction of sentiment reflected in visual content has been presented. To reach this goal a systematic, data-driven methodology was proposed to construct a large-scale Visual Sentiment Ontology (VSO) based on psychology and folksonomies. Further, SentiBank was introduced, a concept detector library of 1,200 ANPs, which is derived from the constructed ontology providing a novel mid-level representation helping to bridge the *affective gap*.

The presented SentiBank features were demonstrated to outperform other low-level feature represen-

tations for the task of sentiment prediction and pornography detection. For the task of detecting emotions reflected in images, SentiBank performed comparable to state-of-the art methods. Finally, all material was released publicly, including the concept ontology, its representing dataset, the detector library, and the benchmark for visual sentiment analysis to the public to stimulate research in this direction.

Considering future work, several exciting questions are open for investigation. First, the cultural influence on perceiving sentiment and emotion is of interest. In particular it would be interesting to investigate if there is a unified understanding of sentiment around the world as done by Darwin [Dar98] for facial expressions in the context of basic emotions. Furthermore, the application of special features such as aesthetic related features used in [MH10] and face expression features offers interesting potential for study. Also, a focus on a refined ontology structure might be beneficial. Additionally, further applications of the proposed work such as advertising, games, augmented vision are imaginable when the cross-domain performance of the detectors is studied in more depth.

Chapter 6

Discussion

People have an inherent need to express themselves. The availability of broadband Internet, low-priced storage, and the omnipresence of camera equipped mobile devices allows us to record, publish, share, and consume digital images and videos without effort. As a consequence multimedia retrieval systems call for new strategies to cope with the increasing scale of current visual databases.

In particular, with respect to social media and multimedia content becoming the dominant content type on the web, *social multimedia* as the fusion of both requires concept detection to move beyond small scale concept vocabularies to be better align with users' information needs. To this end, visual learning of thousands of target concepts must be accomplished to synchronize concept detection with real world events. Furthermore, to provide a comprehensive view of social multimedia, concept detection has to be extended to novel forms of analysis such as the extraction of sentiment reflected in visual content.

Therefore, the core endeavor of this thesis has been to answer the following questions: Can we link concept detection vocabularies to current trends by mining social media? Can we extend concept detection such that a visual learning of thousands of concepts is possible? And, can the notion of semantic concepts be re-thought such that the sentiment reflected in visual content might be extracted?

Starting with the first question, in Chapter 3 dynamic vocabularies for concept detection have been introduced, as being automatically augmented with *trending topics* mined from Google, Wikipedia and Twitter. The selection process was further enhanced by forecasting the progression of trending topics at the very moment they emerge. The presented nearest neighbor sequence matching, which was based on the assumption that semantically similar topics show similar popularity over time, was demonstrated to reliably identify promising trending topics for detector learning. This way trending topics could be added into the detection system either by mapping the trend to a fixed amount of semantic concepts or by a specific training of a trend detector from web video. As a result concept detection was aligned to current real-word events by detecting such trending topics more robustly than with a fixed concept detection vocabulary. Finally, a combination of the proposed marginalization approach with specifically trained trend detectors was shown to further boost recognition performance.

A second key component of the presented concept detection system in this thesis was scalability. This was achieved by utilizing web videos as an alternative source for detector training. This includes the automatic retrieval of training data from platforms like YouTube and the proper handling of user-

generated tags serving as pseudo labels for supervised machine learning. Chapter 4 has investigated the impact of these issues on concept detection, and presented two solutions to deal with them: First, an automatic query construction performing a *concept-to-query mapping* for web video retrieval was presented. This method achieved a retrieval of training data quality comparable with human refined queries. Moreover, to eliminate the remaining effects of pseudo labels, *active relevance filtering* was proposed. It combines automatic relevance filtering with active learning to adopt the statistical models underlying concept detection such that non-relevant content is identified and filtered during detector training. This method was demonstrated to train non-degrading concept detectors with a minimum of user interaction.

Lastly in Chapter 5, concept detection was extended by the notion of *Adjective Noun Pairs* (ANP) as novel entities for concept learning. Such ANPs allow the addition of new levels of differentiation. Instead of detecting of a "dog" concept, the proposed ANPs differentiate between concepts such as "cute dog" or "dangerous dog" and therefore allow to put semantic concepts into context. Founded on psychological theory and driven by an analysis of over 6 Million tags mined from popular image and video sharing platforms, a large scale Visual Sentiment Ontology (VSO) of 3,000 ANPs was constructed. Each ANP from the introduced ontology aims to reflect a strong sentiment, have a link to an emotion, be frequently used, and have a reasonable detection accuracy. Based on this ontology – SentiBank – a detector bank of 1,200 ANP concepts was trained to serve as a mid-level attribute representation of visual content. This effort, as being the first of its kind, was able to provide more robust sentiment predictions from visual content than low-level features and boosted the performance of a combined text-image based sentiment analysis. Additionally, although initially proposed to capture the sentiment being reflected in visual content, this mid-level attribute representation demonstrated its generalization potential to other domains such as emotion detection and digital forensics.

Altogether, the contributions of this thesis can be aligned along the concept detection pipeline as outlined in the framework overview in Figure 1.2. Combining social media analysis and efficient visual learning, the notion of static concept vocabularies has been altered and evolving vocabularies, which are dynamically extended by trending topics, has presented. In parallel the automatic retrieval of web video training data and the refinement of user-generated tags can be considered as a key element of scalable detector training. Finally, SentiBank and its foundation, the VSO, have been suggested to extract another new form of information namely positive, neutral, and negative sentiment. This form of visual recognition itself was made possible by Adjective Noun Pairs, which although treated as semantic concepts have added a new dimension to visual learning.

Appendix A

Fixed Vocabulary Concepts

This section provides an overview of all concepts in the static vocabulary used Chapter 3. These query definition were also used in the following previous publications [UKBB09, Koc11].

Table A.1: A list of all semantic concepts used and the corresponding YouTube queries used to train a static vocabulary concept detection system.

Concept	Query	Category
airplane-flying	airplane & flying -indoor	-
americas-got-talent	americas got talent	-
anime	anime mix	-
aquariums	aquarium fish tank	Animals
arcade	arcade	Travel
asians	asians -hot -sexy -bikini	People
autumn	autumn colors	Travel
baby	baby first	People
badlands	badlands	Travel
balloons	balloons	Entertainm.
baseball	baseball -golf	Sports
basketball	basketball	Sports
beach	beach	Travel
beehive	beehive	Animals
bicycle	bicycle	Vehicles
bikini	bikini	
bill-clinton	bill clinton	News
birds	birds	Animals
blacksmithing	blacksmith	Howto
boat	boat small -rc	Vehicles
boat-ship	ship &(queen|freedom|royal)	Vehicles
boobs	boobs tits	
boxing	boxing	Sports
breakdancing	break dancing	
bridge	bridge -crossing -ship	Travel
brown-bear	brown bear	Animals
bus	bus -van -suv -vw -ride	Vehicles
cake	cake	Howto
camels	camel|dromedar -spider	Animals

Continued on next page

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries used to train the TubeTagger system.

Concept	Query	Category
campus	university campus tour	-
car	car	Vehicles
car-crash	car crash	Vehicles
car-racing	car racing -rc	Sports
cartoon	cartoon	Film
castle	castle &(afar|outside) -inside	Travel
cathedral	cathedral	Travel
cats	cats	Animals
celebration	celebration	Travel
cheerleading	cheerleading	-
choir	choir	-
christmas-tree	christmas tree -fire	-
circus	circus show	-
city-skyline	skyline	Travel
cityscape	cityscape -slideshow -emakina	Travel
classroom	classroom & school -secret	-
clock-tower	clock tower	Travel-
clouds	clouds & beautiful	Travel
cockpit	cockpit -railway -line	Vehicles
commercial	commercial -barack	-
concert	concert	Music
cooking	cooking	Howto
counterstrike-game	counterstrike movie -lego -real	
court	court judge	News
cows	cow	Animals
crane	crane	Vehicles
crash	crash	Vehicles
dam	dam	Travel
dancing	dancing	People
dark-skinned-people	black people	-
darth-vader	darth vader	-
demonstration	protesting	-
desert	desert	Travel
dog	dog	Animals
dogs	dogs	Animals
drawing	drawing	Film
drinking	drinking competition	-
driver	car & vehicle & driver -simulator	-
drummer	drummer	Howto
eiffeltower	eiffeltower	Travel
emergency-vehicle	emergency & vehicle -driver -ride	Vehicles
excavation	excavation	Travel
explosion	explosion	Howto
fence	fence	Travel
fencing	fencing	Sports
ferarri	ferarri	Vehicles
firefighter	firefighter training	-
fireworks	fireworks (nice or beautiful)	-

Continued on next page

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries used to train the TubeTagger system.

Concept	Query	Category
fish	fish	Animals
fishing	fishing	Sports
flood	flood water	News
flower	flower & (bouquet|bloom)	-
food	food delicious	-
football	american football -soccer	Sports
forest	forest	Travel
fountain	fountain	Travel
freeclimbing	freeclimbing	Sports
furniture	furniture	-
garden	garden beautiful -royal -coral	Travel
gardening	gardening	Howto
gas-station	gas station	Travel
georgewbush	george w bush	News
geyser	geyser	Travel
glacier	glacier	Travel
glasses	glasses wearing -not -are	-
golf	golf	Sports
golf-course	golf course flyover	Sports
graffiti	graffiti	-
grand-canyon	grand canyon	Travel
gym	gym	Sports
gymnastics	gymnastics	Sports
hand	hand & daft	-
harbor	harbor & dock	Travel
helicopter	helicopter	Vehicles
highway	highway us route	-
hiking	hiking	Travel
horse	horse	Animals
horse-racing	horse racing	Sports
hospital	hospital & emergency	-
hotel-room	”hotel room”	Travel
house	house sightseeing	Travel
ice-skating	ice skating	Sports
interview	interview	News
iphone	iphone	-
jewellery	jewellery	-
jungle	jungle tropical	Travel
kiss	kissing two	-
kitchen	kitchen -knife -remodel	Howto
laboratory	laboratory tour	-
laundry	laundry	Howto
lava	lava flow	Travel
library	library tour	-
lighthouse	lighthouse	Travel
lightning	lighting strike	Travel
map	map geographic	-
marionette	marionette show	-

Continued on next page

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries used to train the TubeTagger system.

Concept	Query	Category
market	market	Travel
mccain	john mc cain	News
memorial	memorial -day	Travel
military-parade	military parade	-
monitor	screen monitor	-
moon	moon footage	-
mosque	mosque	Travel
motorcycle	(motorcycle or motorbike) -crash	Vehicles
mountain	mountain & panorama	Travel
muppets	muppet show	-
music-video	music video	-
native-american	native american dance	-
neon-sign	neon sign	Travel
nighttime	"by night"	Travel
obama	barrack obama	News
office	office working	-
old-people	"old people"	-
operating-room	operating room	-
orchestra	orchestra symphony	-
origami	origami	Howto
outer-space	universe galaxy -super -song	-
pagoda	pagoda	Travel
parachute	parachute -no	Sports
penguin	penguin	Animals
phone	phone & device	-
piano	piano playing	-
pier	pier	Travel
playground	playground	Travel
poker	poker	Entertainm.
polar-bear	polar bear	Animals
pope	pope benedict	-
pottery	pottery	-
press-conference	press conference	News
procession	procession	Travel
pyramids	pyramid	Travel
race	race	Vehicles
railroad	railroad train -model	Vehicles
rainbow	rainbow beautiful	Travel
rainforest	rain forest	Travel
ranch	ranch	Travel
rc-car	rc car	Vehicles
restaurant	restaurant	Travel
rice-terrace	rice terrace	Travel
riding	horse riding	-
riot	riot	News
river	river	Travel
robot	robot -dance -dancers	-
rocket-launching	rocket launch -model -mini -toy	-

Continued on next page

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries used to train the TubeTagger system.

Concept	Query	Category
rodeo	rodeo bull riding	Sports
rooftop	rooftop	Travel
rugby	rugby	Sports
ruins	ruins -underwater	Travel
runway	runway airport	-
safari	safari	Travel
sailing	sailing	Travel
santa	santa (costume or outfit)	-
secondlife	secondlife	Games
shipwreck	ship wreck	Travel
shooting	shooting gun	-
shopping-mall	shopping (mall or center)	Travel
simpsons	the simpsons homer	-
singing	singing & (gospel|choire)	-
skateboarding	skateboarding	-
skiing	skiing -water	Sports
sky	beautiful sky	Travel
snake	snake	Animals
snooker	snooker	Sports
soccer	soccer	Sports
soldiers	soldiers -child	News
stairs	stairs	Travel
steppe	steppe	Travel
street	street & paved	-
submarine	submarine	Vehicles
subway	subway station	Travel
sunrise	sunrise	Travel
surfing	surfing wave	-
swimming	swimming	Sports
swimming-pools	swimming pool	Travel
sword-fight	sword fight	Sports
talkshow	talkshow	People
tank	tank	Vehicles
tennis	tennis -table	Sports
tent	tent	Travel
themepark	park &(amusement|theme)	Travel
toilet	toilet	-
tony-blair	tony blair	News
tornado	tornado	-
tractor-combine	(harvester or tractor)	Vehicles
traffic	traffic	Travel
traffic-lights	traffic lights	Travel
tunnel	tunnel+ &(through|inside) -approach	Travel
turban	turban	-
two-people	two & people -sleepy -questions	-
underwater	underwater	Travel
us-flag	US flag raised	-
vending-machine	vending machine	Travel

Continued on next page

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries used to train the TubeTagger system.

Concept	Query	Category
videoblog	videoblog	People
waterfall	waterfall	Travel
weather	weather forecast	-
wedding	wedding footage	-
wheel	wheel	Vehicles
windmill	wind mill	Travel
windows-desktop	windows desktop	-
worldofwarcraft	world of warcraft	Entertainm.
wrestling	wrestling	Sports

Appendix B

Visual Sentiment Ontology Structure

This section gives a description of the visual sentiment ontology structure as outlined in Chapter 5. Figure B.1 privides an overview of adjective groups and Figure B.2 provides an overview of the hierarchical noun taxonomy derived from the VSO.

Figure B.1: Groups VSO adjectives. Light pink color indicates the 6 top level nodes and a solid pink node indicates a leaf in the tree structure.

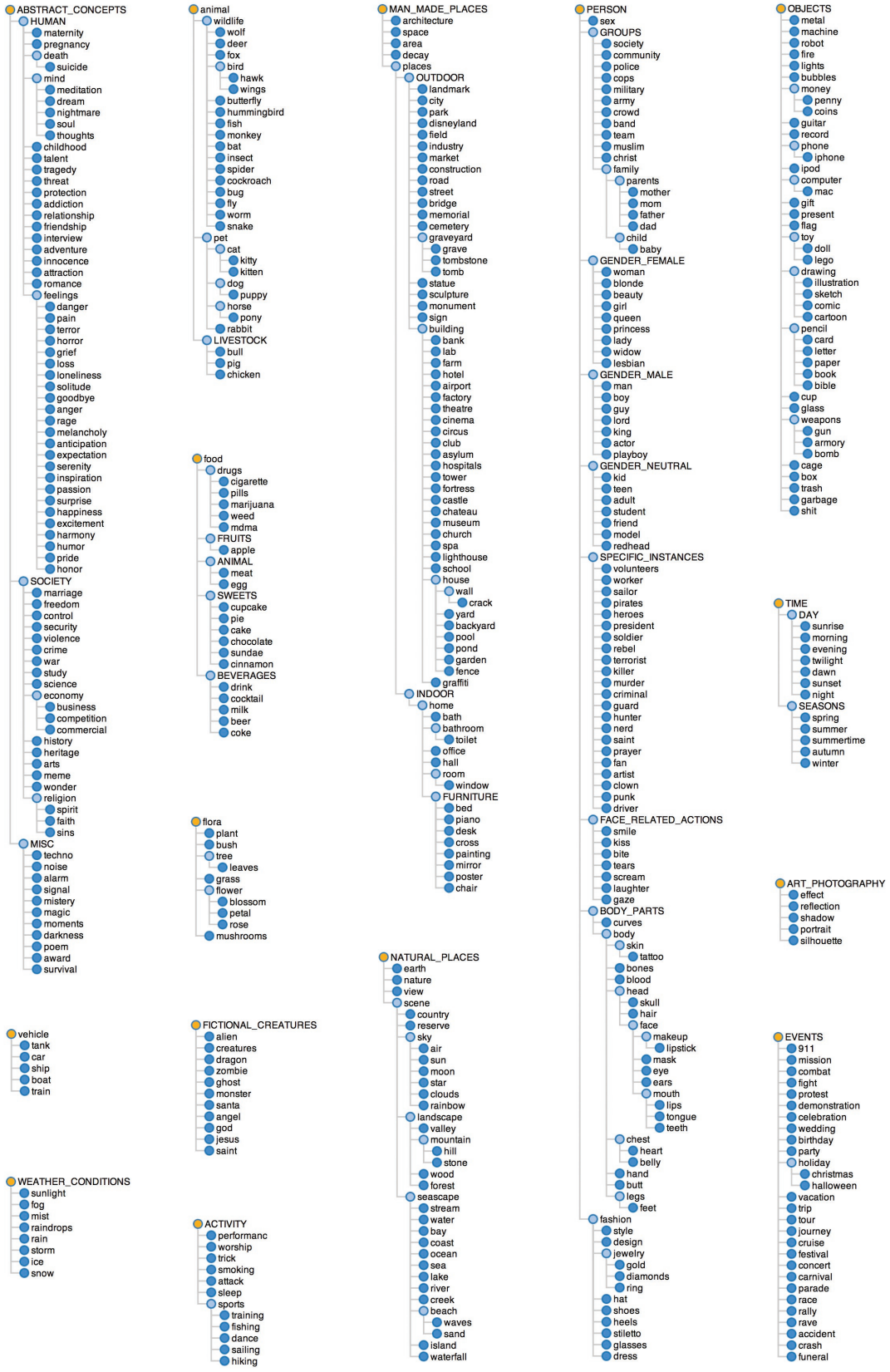


Figure B.2: Hierarchical taxonomy of VSO nouns. Orange color indicates the 15 top level nodes. Light blue color indicates a node with children and a solid blue node indicates a leaf in the tree structure.

Bibliography

[ABC+03] A. Amir, M. Berg, S. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock, J. Smith, B. Tseng, Y. Wu, and D. Zhang. IBM Research TRECVID-2003 Video Retrieval System. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2003.

[ABHD13] T. Althoff, D. Borth, J. Hees, and A. Dengel. Analysis and Forecasting of Trending Topics in Online Media Streams. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 907–916, October 2013.

[ABK+07] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A Nucleus for a Web of Open Data. *Inf. Semantic Web Conference (ISWC)*, pages 722–735, November 2007.

[ACAB99] E. Ardizzone, M. La Cascia, A. Avanzato, and A. Bruna. Video Indexing using MPEG Motion Compensation Vectors. In *Proc. Int. Conf. on Multimedia Computing and Systems*, volume 2, pages 725–729, June 1999.

[ADDK99] Y. Avrithis, A. Doulamis, N. Doulamis, and S. Kollias. A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases. *Computer Vision Image Understanding*, 75(1-2):3–24, 1999.

[All02] J. Allan. Introduction to Topic Detection and Tracking. In *Topic Detection and Tracking*, pages 1–16. Springer, 2002.

[AQ07a] S. Ayache and G. Quénot. Evaluation of Active Learning Strategies for Video Indexing. *Signal Processing: Image Communication*, 22(7-8):692–704, 2007.

[AQ07b] S. Ayache and G. Quénot. TRECVID 2007 Collaborative Annotation using Active Learning. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2007.

[AQ08] S. Ayache and G. Quénot. Video Corpus Annotation using Active Learning. In *Proc. European Conf. on Information Retrieval (ECIR)*, April 2008.

[ASD12] T. Althoff, H. Song, and T. Darrell. Detection Bank: An Object Detection-based Video Representation for Multimedia Event Recognition. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1065–1068, October 2012.

[ATY09] H Aradhye, G. Toderici, and J. Yagnik. Video2Text: Learning to Annotate Video Content. In *Proc. IEEE ICDM Workshop on Internet Multimedia Mining*, pages 144–151, December 2009.

[AWBG07] E. Adar, D. Weld, B. Bershad, and S. Gribble. Why we Search: Visualizing and Predicting User Behavior. In *Proc. Int. Conf. World Wide Web (WWW)*, pages 161–170, May 2007.

[BA96] M. Black and P. Anandan. The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields. *Computer Vision and Image Understanding.*, 63(1):75–104, 1996.

[Bad11] D. Badawi. Audio Features for Automatic Video Tagging. Technical report, Bachelor Thesis, German University in Cairo, Egypt, 2011.

[BAH12] R. Bandari, S. Asur, and B. Huberman. The Pulse of News in Social Media: Forecasting Popularity. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, pages 26–33, June 2012.

[BB96] S. Beauchemin and J. Barron. The Computation of Optical Flow. *ACM Computing Surveys*, 27(3):433–467, 1996.

[BBC06] Britain is 'Surveillance Society'. in BBC News; available from `http://news.bbc.co.uk/1/hi/uk/6108496.stm` (retrieved: Feb'09), November 2006.

[BBDB+10] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, and G. Serra. Tag Suggestion and Localization in User-generated Videos based on Social Knowledge. In *Proc. ACM Workshop on Social Media*, pages 3–8, October 2010.

[BBS10] T. Berg, A. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *Proc. European. Conf. on Computer Vision (ECCV)*, pages 663–676. Springer, September 2010.

[BCC+13] L. Brown, L. Cao, S.-F. Chang, Y. Cheng, A. Choudhary, N. Codella, C. Cotton, D. Ellis, Q. Fan, R. Feris, et al. IBM Research and Columbia University TRECVID-2013 Multimedia Event Detection (MED), Multimedia Event Recounting (MER), Surveillance Event Detection (SED), and Semantic Indexing (SIN) Systems. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2013.

[BF06] T. Berg and D. Forsyth. Animals on the Web. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1463–1470, June 2006.

[BHK+09] D. Borth, J. Hees, M. Koch, A. Ulges, C. Schulze, T. Breuel, and R. Paredes. TubeFiler – an Automatic Web Video Categorizer. In *Proc. ACM Int. Conf. on Multimedia (ACM MM), Multimedia Grand Challenge*, pages 1111–1112, October 2009.

[Bis07] C. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2007.

[BJC+13] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 223–232, October 2013.

[BJCC13] D. Borth, R. Ji, T. Chen, and S.-F. Chang. SentiBank: Large-Scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content. In *Proc. ACM Int. Conf. on Multimedia (ACM MM), Demo Session*, pages 459–460, October 2013.

[BKUB09] D. Borth, M. Koch, A. Ulges, and T. Breuel. DFKI-IUPR Participation in the TRECVID'09 High-level Feature Extraction Task. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2009.

[BL12] D. Borth and M. Löffler. Making Web Videos more Accessible to Advertising: The Winner of McKinsey's Business Technology Award. *McKinsey Quarterly, Business Technology*, 23:40–44, 2012.

[BLS01] T. Bozios, G. Lekakos, and V. Skoularidou. Advanced Techniques for Personalized Advertising in a Digital TV Environment: The IMedia System. In *Proc. of the eBusiness and eWork Conference*, 2001.

[BMZ11] J. Bollen, H. Mao, and X. Zeng. Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1):1–8, 2011.

[BNG11] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics: Real-world Event Identification on Twitter. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, July 2011.

[Bob01] M Bober. MPEG-7 Visual Shape Descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6):716–719, 2001.

[BSS11] S. Bhattacharya, R. Sukthankar, and M. Shah. A Holistic Approach to Aesthetic Enhancement of Photographs. *ACM Trans. Multimedia Computing, Communications, and Applications (TOMCCAP)*, 7(1):21, 2011.

[BSUB08] D. Borth, C. Schulze, A. Ulges, and T. Breuel. Navidgator - Similarity Based Browsing for Image & Video Databases. In *Proc. KI 2008*, pages 22–29, September 2008.

[BTvG06] H. Bay, T. Tuytelaars, and L. van Gool. SURF: Speeded Up Robust Features. In *Proc. European. Conf. on Computer Vision (ECCV)*, pages 404–417, May 2006.

[BUB10] D. Borth, A. Ulges, and T. Breuel. Relevance Filtering meets Active Learning: Improving Web-based Concept Detectors. In *Proc. ACM Int. Conf. on Multimedia Information Retrieval (ACM MIR)*, pages 25–34, March 2010.

[BUB11a] D. Borth, A. Ulges, and T. Breuel. Automatic Concept-to-Query Mapping for Web-based Concept Detector Training. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1453–1456, November 2011.

[BUB11b] D. Borth, A. Ulges, and T.M. Breuel. Lookapp - Interactive Construction of Web-based Concept Detectors. In *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR)*, pages 66–68, April 2011.

[BUB12] D. Borth, A. Ulges, and T. Breuel. Dynamic Vocabularies for web-based Concept Detection by Trend Discovery. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 977–980, October 2012.

[BUSB08] D. Borth, A. Ulges, C. Schulze, and T. Breuel. Keyframe Extraction for Video Tagging and Summarization. In *Proc. Informatiktage 2008*, pages 45–48, March 2008.

[BYRN+99] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern Information Retrieval*. ACM press New York, 1999.

[BZM07] A. Bosch, A. Zisserman, and X. Muoz. Image Classification using Random Forests and Ferns. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 1–8, 2007.

[C. 09] C. Snoek et al. The MediaMill TRECVID 2009 Semantic Video Search Engine. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2009.

[Can86] J. Canny. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[CBP09] S. Choudhury, J. G. Breslin, and A. Passant. Enrichment and Ranking of the YouTube Tag Space and Integration with the Linked Data Cloud. In *Proc. Int. Semantic Web Conference (ISWC)*, pages 747–762, 2009.

[CCC+11] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu amd A. Natsev, and J. Smith. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, December 2011.

[CCHW05] M. Chen, M. Christel, A. Hauptmann, and H. Wactlar. Putting Active Learning into Multimedia Applications: Dynamic Definition and Refinement of Concept Classifiers. In *Proc. Int. Conf. on Multimedia (ACM MM)*, pages 902–911, November 2005.

[CCMV07] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[CCS+10] Z. Chen, J. Cao, Y. Song, Y. Zhang, and J. Li. Web Video Categorization based on Wikipedia Categories and Content-duplicated Open Resources. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1107–1110, October 2010.

[CDBP99] J. Corridoni, A. Del Bimbo, and P. Pala. Image Retrieval by Color Semantics. *Multimedia Systems*, 7(3):175–183, 1999.

[CEJ+06] M. Campbell, S. Ebadollahi, D. Joshi, M. Naphade, A. Natsev, J. Seidl, J. Smith, K. Scheinberg, J. Tesic, and L. Xie. IBM Research TRECVID-2006 Video Retrieval System. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2006.

[CEJ+07] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, and J. Luo. Large-scale Multimodal Semantic Concept Detection for Consumer Video. In *Proc. Int. Workshop Multimedia Information Retrieval (MIR)*, pages 255–264, September 2007.

[CH05] M. Christel and A. Hauptmann. The Use and Utility of High-Level Semantic Features in Video Retrieval. In *Proc. ACM Int. Conf. Image and Video Retrieval (CIVR)*, pages 134–144, July 2005.

[CH09] M. Clements and D. Hendry. Forecasting Annual UK Inflation using an Econometric Model over 1875–1991. *Frontiers of Economics and Globalization*, 3:3–39, 2009. Forecasting in the Presence of Structural Breaks and Model Uncertainty.

[CHL+07] M. Campbell, A. Haubold, M. Liu, A. Natsev, J. Smith, J. Tesic, L. Xie, R. Yan, and J. Yang. IBM Research TRECVID-2007 Video Retrieval System. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2007.

[CL01] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[CL09] M. Cheong and V. Lee. Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base. In *Proc. ACM Workshop on Social Web Search and Mining (SWSM)*, pages 1–8, November 2009.

[CLL+06] J. Cao, Y. Lan, J. Li, Q. Li, X. Li, F. Lin, X. Liu, L. Luo, W. Peng, D. Wang, H. Wang, Z.Wang, Z. Xiang, J. Yuan, W. Zheng, B. Zhang, J. Zhang, L. Zhang, and X. Zhang. Intelligent Multimedia Group of Tsinghua University at TRECVID 2006. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2006.

[CLVZ11] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The Devil is in the Details: An Evaluation of Recent Feature Encoding Methods. In *ePrints Pascal Network*, 2011.

[CMC05] S.-F. Chang, R. Manmatha, and T.-S. Chua. Combining Text and Audio-visual Features in Video Indexing. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'05)*, volume 5, pages v–1005, 2005.

[CNL+04] T.S. Chua, S.Y. Neo, K.Y. Li, G. Wang, R. Shi, M. Zhao, H. Xu, S. Gao, and T.L. Nwe. TRECVID 2004 Search and Feature Extraction Task by NUS PRIS. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, 2004.

[CSBB97] S.-F. Chang, J. Smith, M. Beigi, and A. Benitez. Visual Information Retrieval from Large Distributed Online Repositories. *Communications of the ACM*, 40(12):63–71, 1997.

[CSG+02] C. Cieri, S. Strassel, D. Graff, N. Martey, K. Rennert, and M. Liberman. Corpora for Topic Detection and Tracking. *Topic Detection and Tracking*, pages 33–66, 2002.

[CSK11] A. Criminisi, J. Shotton, and E. Konukoglu. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-supervised Learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, 5(6):12, 2011.

[CTGC05] E.Y. Chang, S. Tong, KS Goh, and C.W. Chang. Support Vector Machine Concept-Dependent Active Learning for Image Retrieval. *IEEE Trans. on Multimedia*, 2, 2005.

[CV12] H. Choi and H. Varian. Predicting the Present with Google Trends. *Economic Record*, 88(s1):2–9, 2012.

[Dar98] C. Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA, 1872 / 1998.

[DC01] S. Das and M. Chen. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. In *Proc. Annual Meeting of European Finance Association*, 2001.

[DCCC11] M. De Choudhury, S. Counts, and M. Czerwinski. Find Me the Right Content! Diversity-Based Sampling of Social Media Spaces for Topic-Centric Search. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, July 2011.

[DD03] D. Dementhon and D. Doermann. Video Retrieval using Spatio-Temporal Descriptors. In *Proc. Int. Conf. on Multimedia (ACM MM)*, pages 508–517, October 2003.

[DDF+90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[DDS+09] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, July 2009.

[DGS11] E. Dan-Glauser and K. Scherer. The Geneva Affective Picture Database (GAPED): a New 730-picture Database Focusing on Valence and Normative Significance. *Behavior Research Methods*, 43(2):468–477, 2011.

[DHS00] R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[Die00] T. Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine learning*, 40(2):139–157, 2000.

[dJGHN99] F. de Jong, J.-L. Gauvain, J. Hartog, and K. Netter. OLIVE: Speech-based Video Retrieval. In *Proc. European Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 75–80. IRIT, 1999.

[DJLW06] R. Datta, D. Joshi, J. Li, and J. Wang. Studying Aesthetics in Photographic Images using a Computational Approach. In *Proc. European. Conf. on Computer Vision (ECCV)*, pages 288–301. Springer, May 2006.

[DJLW07] R. Datta, D. Joshi, J. Li, and J. Wang. Tagging over Time: Real-world Image Annotation by Lightweight Meta-Learning. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 393–402, 2007.

[DJLW08] R. Datta, D. Joshi, J. Li, and J. Wang. Image Retrieval: Ideas, Influences, and Trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[DKN05] T. Deselaers, D. Keysers, and H. Ney. Discriminative Training for Object Recognition using Image Patches. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 157–162, June 2005.

[DKN08] T. Deselaers, D. Keysers, and H. Ney. Features for Image Retrieval: An Experimental Comparison. *Information Retrieval*, 11:77–107, 2008.

[DLL+10] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, et al. The YouTube Video Recommendation System. In *Proc. ACM Int. Conf. on Recommender Systems (RecSys)*, pages 293–296, January 2010.

[DLP03] K. Dave, S. Lawrence, and D. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proc. ACM Int. Conf. on World Wide Web (WWW)*, pages 519–528, May 2003.

[DLR77] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[DLW05] R. Datta, J. Li, and J. Wang. Content-based Image Retrieval: Approaches and Trends of the New Age. In *Proc. Int. Workshop on Multimedia Information Retrieval (MIR)*, pages 253–262, October 2005.

[DPK+12] S. Diplaris, S. Papadopoulos, I. Kompatsiaris, A. Goker, A. Macfarlane, J. Spangenberg, H. Hacid, L. Maknavicius, and M. Klusch. SocialSensor: Sensing User Generated Input for Improved Media Discovery and Experience. In *Proc. ACM Int. Conf. on World Wide Web (WWW)*, pages 243–246, April 2012.

[DPN08] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-Visual-Words Models for Adult Image Classification and Filtering. In *Proc. Int. Conf. Pattern Recognition (ICPR)*, pages 1–4, December 2008.

[dRSW08] O. de Rooij, C.G.M. Snoek, and M. Worring. MediaMill: Fast and Effective Video Search using the Forkbrowser. In *Proc. Int. Conf. on Content-based Image and Video Retrieval (CIVR)*, pages 561–562, 2008.

[DSDVL02] A. Datta, M. Shah, and N. Da Vitoria Lobo. Person-on-person Violence Detection in Video Data. In *Int. Conf. on Pattern Recognition (ICPR)*, volume 1, pages 433–438, 2002.

[DT05] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, June 2005.

[DUBW09] M. Duan, A. Ulges, T. Breuel, and X.-Q. Wu. Style Modeling for Tagging Personal Photo Collections. In *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, pages 1–8, July 2009.

[DZS+02] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of Video-content Analysis and Retrieval. *IEEE MultiMedia*, 9(3):42–55, 2002.

[E+93] P. Ekman et al. Facial Expression and Emotion. *American Psychologist*, 48:384–384, 1993.

[ea11] C. Snoek et al. The MediaMill TRECVID 2011 Semantic Video Search Engine. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, 2011.

[Ege97] M Egenhofer. Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages & Computing*, 8(4):403–424, 1997.

[ES06] A. Esuli and F. Sebastiani. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proc. Language Resources and Evaluation Conference (LREC)*, volume 6, pages 417–422, 2006.

[EVGW+08] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, October 2008.

[EVGW+10] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Int. Journal of Computer Vision*, 88(2):303–338, June 2010.

[FB04] F. Fraundorfer and H. Bischof. Evaluation of Local Detectors on Non-Planar Scenes. In *Annual Workshop of the Austrian Association for Pattern Recognition (OAGM/AAPR)*, pages 125–132, June 2004.

[FEHF09] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785, June 2009.

[Fel98] C. Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[FFFPZ05] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. *Computer Vision*, 2:1816–1823, 2005.

[FFP05] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, June 2005.

[FML04] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1002–1009, June 2004.

[FRSRAF99] F. Fernández-Rodrıguez, S. Sosvilla-Rivero, and J. Andrada-Félix. Exchange-rate Forecasts with Simultaneous Nearest-Neighbour Methods: Evidence from the EMS. *International Journal of Forecasting*, 15(4):383–392, 1999.

[FSN+95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, et al. Query by Image and Video Content: The QBIC System. *IEEE Computer*, 28(9):23–32, 1995.

[FZ07] V. Ferrari and A. Zisserman. Learning Visual Attributes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, December 2007.

[GH06] S.A. Golder and B.A. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198, 2006.

[GHL+10] S. Goel, J. Hofman, S. Lahaie, D. Pennock, and D. Watts. Predicting Consumer Behavior with Web Search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, 2010.

[GHT04] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated Trend Discovery for Weblogs. In *Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[GMH+08] Z. Gu, T. Mei, X.S. Hua, J. Tang, and X. Wu. Multi-layer Multi-Instance Learning for Video Concept Detection. *IEEE Trans. on Multimedia*, 10(8):1605–1616, 2008.

[Gon08] González-Díaz, I. et al. UC3M High Level Feature Extraction at TRECVID 2008. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2008.

[GPK02] S. Grigorescu, N. Petkov, and P. Kruizinga. Comparison of Texture Features based on Gabor Filters. *IEEE Trans. on Image Processing*, 11(10):1160–1167, 2002.

[GY08] U. Gargi and J. Yagnik. Solving the Label Resolution Problem in Supervised Video Content Classification. In *Proc. ACM Int. Conf. on Multimedia Information Retrieval (ACM MIR)*, pages 276–282, October 2008.

[Han02] A. Hanjalic. Shot-boundary Detection: Unraveled and Resolved? *IEEE Trans. Circuits and Systems for Video Technology*, 12(2):90–105, 2002.

[Hau05] A. Hauptmann. Lessons for the Future from a Decade of Informedia Video Analysis Research. In *Int. Conf. on Image and Video Retrieval (CIVR)*, pages 1–10. Springer, July 2005.

[HBDP12] K. Hermann, P. Blunsom, C. Dyer, and S. Pulman. Learning Semantics and Selectional Preference of Adjective-Noun Pairs. In *Proc. ACL Int. Conf. on Lexical and Computational Semantics-Volume*, pages 70–74, June 2012.

[HBU12] D. Henter, D. Borth, and A. Ulges. Tag Suggestion on Youtube by Personalizing Content-based Auto-Annotation. In *Proc. ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, pages 41–46, October 2012.

[HK08] R. Hyndman and Y. Khandakar. Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software*, 27(3), 2008.

[HKC06] W. Hsu, L. Kennedy, and S.-F. Chang. Video Search Reranking via Information Bottleneck Principle. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 35–44, October 2006.

[HKC07] W. Hsu, L. Kennedy, and S.-F. Chang. Reranking Methods for Visual Search. *IEEE Multimedia*, 14(3):14–22, 2007.

[HKL12] A. Hanjalic, C. Kofler, and M. Larson. Intent and its Discontents: The User at the Wheel of the Online Video Search Engine. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1239–1248, October 2012.

[HKM+97] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image Indexing using Color Correlograms. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 762–768, June 1997.

[HL04] S. Helmer and D. Lowe. Object Class Recognition with Many Local Features. In *Proc. Int. Conf. Computer Vision and Pattern Recognition Workshop*, pages 187–195, June 2004.

[HLB99] A. Hanjalic, R. Lagendijk, and J. Biemond. Automated High-level Movie Segmentation for Advanced Video-retrieval Systems. *IEEE Trans. on Circuits and Systems for Video Technology*, 9(4):580–588, 1999.

[HLY+06] A. Hauptmann, W. Lin, R. Yan, J. Yang, and M. Chen. Extreme Video Retrieval: Joint Maximization of Human and Computer Performance. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 385–394, 2006.

[HM00] R. Hammoud and R. Mohr. A Probabilistic Framework of Selecting Effective Key Frames for Video Browsing and Indexing. In *Proc. Int. Workshop on Real-Time Image Sequence Analysis*, pages 79–88, August 2000.

[HN07] A. Haubold and M. Naphade. Classification of Video Events using 4-dimensional Time-Compressed Motion Features. In *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, pages 178–185, July 2007.

[HOdJ07] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of Heterogeneous Multimedia Content using Automatic Speech Recognition. In *Proc. Int. Conf. Semantics and Digital Media Technology*, pages 78–90, December 2007.

[HS88] C. Harris and M. Stevens. A Combined Corner and Edge Detector. In *Proc. Alvey Vision Conference*, pages 147–151, May 1988.

[HSdRS12] B. Huurnink, C. Snoek, M. de Rijke, and A. Smeulders. Content-based Analysis Improves Audiovisual Archive Retrieval. *IEEE Trans. on Multimedia*, 14(4):1166–1178, 2012.

[HSSV03] E. Hyvönen, S. Saarela, A. Styrman, and K. Viljanen. Ontology-based image retrieval. In *Proc. ACM Int. Conf. on World Wide Web (WWW)*, 2003.

[HvdSS13] A. Habibian, K. van de Sande, and C. Snoek. Recommendations for Video Event Recognition using Concept Vocabularies. In *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR)*, pages 89–96, April 2013.

[HYL07] A. Hauptmann, R. Yan, and W. Lin. How many High-Level Concepts will Fill the Semantic Gap in News Video Retrieval? In *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, pages 627–634, July 2007.

[HZ99] A. Hanjalic and H. Zhang. An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis. *IEEE Trans. Circuits and Systems for Video Technology*, 9(8):1280–1289, 1999.

[Inc12] Cisco Systems Inc. Cisco Visual Networking Index: Forecast and Methodology, 2011-2016. available from `http://www.cisco.com` (retrieved: February'13), February 2012.

[IXTO11] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What Makes an Image Memorable? In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2011.

[JDF+11] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and Emotions in Images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.

[JGC+10] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The Wisdom of Social Multimedia: Using Flickr for Prediction and Forecast. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1235–1244, October 2010.

[JH94] R. Jain and A. Hampapur. Metadata in Video Databases. *ACM Sigmod Record*, 23(4):27–33, 1994.

[JM04] J. Jeon and R. Manmatha. Using Maximum Entropy for Automatic Image Tagging. In *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, pages 24–32, July 2004.

[JNC09] Y.G. Jiang, C.W. Ngo, and S.F. Chang. Semantic Context Transfer across Heterogeneous Sources for Domain Adaptive Video Search. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, October 2009.

[JNY07] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, pages 494–501, July 2007.

[JS04] F. Jurie and C. Schmid. Scale-invariant Shape Features for Recognition of Object Categories. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 90–96, June 2004.

[JUB09] C. Jansohn, A. Ulges, and T. Breuel. Detecting Pornographic Video Content by Combining Image Features with Motion Information. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 601–604, October 2009.

[JWW+12] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. Can we Understand van Gogh's Mood?: Learning to Infer Affects from Images in Social Networks. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 857–860, October 2012.

[JYC+11] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. Loui. Consumer Video Understanding: A Benchmark Database and An Evaluation of Human and Machine Performance. In *Proc. Int. Conf. on Multimedia Retrieval (ICMR)*, page 29, April 2011.

[JYCN08] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection. Technical report, Columbia University, 2008.

[JZSC09] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188, 2009.

[KB01] T. Kadir and M. Brady. Saliency, Scale and Image Description. *Int. Journal Computer Vision*, 45(2):83–105, 2001.

[KBBN09] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and Simile Classifiers for Face Verification. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 365–372, 2009.

[KCK06] L. Kennedy, S.-F. Chang, and I. Kozintsev. To Search or to Label?: Predicting the Performance of Search-based Automatic Image Classifiers. In *Workshop Multimedia Information Retrieval*, 2006.

[KEG+08] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems: An evaluation. *ACM Transactions on the Web (TWEB)*, 2(4):1–34, 2008.

[KGZC13] E. Konukoglu, B. Glocker, D. Zikic, and A. Criminisi. Neighbourhood Approximation using Randomized Forests. *Medical Image Analysis*, 17(7):790–804, 2013.

[KHN+06] L. Kennedy, A. Hauptmann, M. Naphade, J. Smith, and S.-F. Chang. LSCOM Lexicon Definitions and Annotations Version 1.0. Technical report, Columbia University, 2006.

[KL09] C. Kofler and M. Lux. An Exploratory Study on the Explicitness of User Intentions in Digital Photo Retrieval. In *Proc. Int. Conf. on Knowledge Management and Knowledge Technologies (I-KNOW)*, pages 208–214, September 2009.

[KLPM10] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proc. ACM Int. Conf. on World Wide Web (WWW)*, April 2010.

[KLS13] S. Kordumova, X. Li, and C. Snoek. Evaluating Sources and Strategies for Learning Video Concepts from Social Media. In *Proc. IEEE Int. Workshop on Content-Based Multimedia Indexing*, pages 91–96, June 2013.

[KNC05] L. Kennedy, A. Natsev, and S.-F. Chang. Automatic Discovery of Query-Class-dependent Models for Multimodal Search. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 882–891, October 2005.

[Koc11] M. Koch. Ad Targeting for Web Video by Automatic Video Annotation. Technical report, Master Thesis, University of Kaiserslautern (available from http://madm.dfki.de/teaching), 2011.

[KSH12] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, December 2012.

[LBC99] P. Lang, M. Bradley, and B. Cuthbert. International Affective Picture System (IAPS): Technical Manual and Affective Ratings, 1999.

[LDK00] H. Li, D. Doermann, and O. Kia. Automatic Text Detection and Tracking in Digital Video. *IEEE Trans. Image Processing*, 9(1):147–156, 2000.

[Lew98] D. Lewis. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proc. Europ. Conf. Machine Learning (ECML)*, pages 4–15, April 1998.

[LFKS09] H. Luo, J. Fan, D. Keim, and S. Satoh. Personalized News Video Recommendation. In *Advances in Multimedia Modeling*, pages 459–471. Springer, 2009.

[LFXH12] B. Li, S. Feng, W. Xiong, and W. Hu. Scaring or Pleasing: Exploit Emotional Impact of an Image. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1365–1366, October 2012.

[LG94] D. Lewis and W. Gale. A Sequential Algorithm for Training Text Classifiers. In *Proc. Int. Conf. Research and Development in Information Retrieval*, pages 3–12, July 1994.

[LGS+11] X. Li, E. Gavves, C. Snoek, M. Worring, and A. Smeulders. Personalizing Automated Image Annotation using Cross-Entropy. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 233–242, November 2011.

[LH10] X. Liu and B. Huet. Concept Detector Refinement using Social Videos. In *Proc. Int. Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval*, pages 19–24, October 2010.

[LHY+09] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag Ranking. In *Proc. Int. Conf. on World Wide Web (WWW)*, pages 351–360, April 2009.

[Lie99] R. Lienhart. Comparison of Automatic Shot Boundary Detection Algorithms. In *Storage and Retrieval for Image and Video Databases*, pages 290–301, January 1999.

[Lie01] R. Lienhart. Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. *Int. Journal of Image and Graphics*, 1(3):469–286, 2001.

[Lin98] T. Lindeberg. Feature Detection with Automatic Scale Selection. *Int. Journal Computer Vision*, 30(2):77–116, 1998.

[LLZ+11] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale Image Classification: Fast Feature Extraction and SVM Training. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2011.

[LNH09] C. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes by Between-class Attribute Transfer. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958, June 2009.

[LNJ10] M. Larson, E. Newman, and G. Jones. Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 354–368. Springer, 2010.

[Low99] D. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 1150–1157, September 1999.

[Low04] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. Journal Computer Vision*, 60(2):91–110, 2004.

[LSE+12] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, and G. Jones. The Community and the Crowd: Multimedia Benchmark Dataset Development. *IEEE Multimedia*, 19:15–23, 2012.

[LSFFX10] L.-J. Li, H. Su, L. Fei-Fei, and E. Xing. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1378–1386, December 2010.

[LSP06] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, June 2006.

[LSW08] X. Li, C. Snoek, and M. Worring. Learning Tag Relevance by Neighbor Voting for Social Image Retrieval. In *Proc. ACM Int. Conf. on Multimedia Information Retrieval (ACM MIR)*, pages 180–187, 2008.

[LSW09] X. Li, C. Snoek, and M. Worring. Learning Social Tag Relevance by Neighbor Voting. *IEEE Trans. on Multimedia*, 11(7):1310–1322, November 2009.

[LSW10] X. Li, C. Snoek, and M. Worring. Unsupervised Multi-Feature Tag Relevance Learning for Social Image Retrieval. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pages 10–17, July 2010.

[LSWS12] X. Li, C. Snoek, M. Worring, and A. Smeulders. Harvesting Social Images for Bi-Concept Search. *IEEE Trans. Multimedia*, 14(4):1091–1104, 2012.

[LW09] J. Lin and W. Wang. Weakly-supervised Violence Detection in Movies with Audio and Video based Co-Training. In *Advances in Multimedia Information Processing*, pages 930–935. Springer, December 2009.

[LWFF07] L.-J. Li, G. Wang, and L. Fei-Fei. OPTIMOL: Automatic Object Picture Collection via Incremental Model Learning. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 57–64, June 2007.

[Mar04] J. Martinez. MPEG-7 Overview (Version 10), October 2004.

[MCMP02] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *Proc. British Machine Vision Conference (BMVC)*, pages 384–393, September 2002.

[MGP04] F. Monay and D. Gatica-Perez. PLSA-based Image Annotation: Constraining the Latent Space. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 348–351, October 2004.

[MGvdSS13a] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek. Searching Informative Concept Banks for Video Event Detection. In *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR)*, pages 255–262, April 2013.

[MGvdSS13b] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek. Searching Informative Concept Banks for Video Event Detection. In *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR)*, pages 255–262, April 2013.

[MH10] J. Machajdik and A. Hanbury. Affective Image Classification using Features Inspired by Psychology and Art Theory. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 83–92, October 2010.

[MHL08] T. Mei, X.-S. Hua, and S. Li. Contextual In-Image Advertising. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 439–448, October 2008.

[MHL09] Tao Mei, Xian-Sheng Hua, and Shipeng Li. VideoSense: A Contextual In-video Advertising System. *IEEE Trans. Circuits and Syst. for Video Technology*, 19:1866–1879, December 2009.

[MHS13] M. Mazloom, A. Habibian, and C. Snoek. Querying for Video Events by Semantic Signatures from Few Examples. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 609–612, October 2013.

[Mik03] K. Mikolajczyk. A Performance Evaluation of Local Descriptors. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 257–263, June 2003.

[Mil95] G. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[MLS06] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple Object Class Detection with a Generative Model. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 26–36, June 2006.

[MM96] B. Manjunath and W.-Y. Ma. Texture features for Browsing and Retrieval of Image Data. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[MOVY01] B. Manjunath, J.-R. Ohm, V. Vasuvedan, and A. Yamada. Color and Texture Descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[MPLC11] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the Aesthetic Quality of Photographs using Generic Image Descriptors. In *Proc. Int. Conf. on Computer Vision (ICCV)*, November 2011.

[MRY06] P. Mundur, Y. Rao, and Y. Yesha. Keyframe-based Video Summarization using Delaunay Clustering. *Int. Journal Digital Libraries*, 6(2):219–232, 2006.

[MS04] K. Mikolajczyk and C. Schmid. Scale and Affine Invariant Interest Point Detectors. *Int. Journal Computer Vision*, 60(1):63–86, 2004.

[MZ03] Y. Ma and H. Zhang. Motion Pattern-based Video Classification and Retrieval. *EURASIP Journal on Advances in Signal Processing*, 2003(1):199–208, 2003.

[NBE+93] W. Niblack, R. Barber, Q. Equitz, M. Fickner, E. Glasman, D. Petkovic, and P. Yanker. The QBIC Project: Querying Images by Content using Color, Texture and Shape. In *SPIE Conf. on Geometric Methods in Computer Vision*, pages 23–32, 1993.

[NBS+02] M. Naphade, S. Basu, J. Smith, C.-Y. Lin, and B. Tseng. Modeling Semantic Concepts to Support Query by Keywords in Video. In *Proc. IEEE Int Conf on Image Processing (ICIP)*, pages 145–148, September 2002.

[NH01] M. Naphade and T. Huang. A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval. *IEEE Trans. Multimedia*, 3(1):141–151, 2001.

[NHT+07] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic Concept-based Query Expansion and Re-ranking for Multimedia Retrieval. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 991–1000, September 2007.

[NJW+09] C.-W. Ngo, Y.-G. Jiang, X.-Y. Wei, W. Liu, S. Zhu, and S.-F. Chang. VIREO/DVMM at TRECVID 2009: High-Level Feature Extraction, Automatic Video Search, and Conctent-based Copy Detection. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2009.

[NKK+05] M. Naphade, L. Kennedy, J. Kender, S.-F. Chang, J. Smith, P. Over, and A. Hauptmann. A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. Technical report, IBM Research Division, 2005.

[NLM99] K. Nigam, J. Lafferty, and A. Mccallum. Using Maximum Entropy for Text Classification. In *Proc. IJCAI Workshop Machine Learning for Information Filtering*, pages 61–67, July 1999.

[NMP10] G. Mann N. Morsillo and C. Pal. YouTube Scale, Large Vocabulary Video Annotation. In *Video Search and Mining*. Springer, 2010.

[NS04a] M. Naphade and J. Smith. On the Detection of Semantic Concepts at TRECVID. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 660–667, December 2004.

[NS04b] H.T. Nguyen and A. Smeulders. Active learning using Pre-Clustering. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 623–630, July 2004.

[NST+06] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[NTYS07] A. Natsev, J. Tešić, R. Yan, and J. Smith. IBM Multimedia Search and Retrieval System. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pages 645–645, July 2007.

[NY03] T. Nasukawa and J. Yi. Sentiment Analysis: Capturing Favorability using Natural Language Processing. In *Proc. ACM Int. Conf. on Knowledge Capture*, pages 70–77, October 2003.

[NZKC06] S.Y. Neo, J. Zhao, M.Y. Kan, and T.S. Chua. Video Retrieval using High-Level Features: Exploiting Query Matching and Confidence-based Weighting. *Proc. Int. Conf. on Image and Video Retrieval (CIVR)*, pages 143–152, 2006.

[OAF+10] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. Quénot. TRECVID 2010–An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. In *Proc. NIST TRECVID Workshop*, November 2010.

[OAM+11] P. Over, G. Awad, M. Michael, J.Fiscus, W. Kraaij, and A. Smeaton. TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. NIST TRECVID Workshop*, December 2011.

[OAM+12] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. Smeaton, and G. Quénot. TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2012*, November 2012.

[OAM+13] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. Smeaton, and G. Quénot. TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. NIST TRECVID 2013*, November 2013.

[OAR+09] P. Over, G. Awad, T. Rose, J. Fiscus, W. Kraaij, and A. Smeaton. TRECVID 2009 – Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. NIST TRECVID Workshop*, November 2009.

[OBRS10] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, pages 122–129, May 2010.

[O'C91] B. O'Connor. Selecting Key Frames of Moving Image Documents: A Digital Environment for Analysis and Navigation. *Microcomputers for Information Management*, 8(2):119–33, 1991.

[OPM+12] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no More: First Story Detection using Twitter and Wikipedia. In *Proc. Int. Workshop on Time-aware Information Access (TAIA)*, 2012.

[OST57] C. Osgood, G. Suci, and P. Tannenbaum. *The Measurement of Meaning*, volume 47. Urbana: University of Illinois Press, 1957.

[OT01] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. Journal of Computer Vision*, 42(3):145–175, 2001.

[Pai63] A. Paivio. Learning of Adjective-Noun Paired Associates as a Function of Adjective-Noun Word Order and Noun Abstractness. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 17(4):370, 1963.

[PCI+07] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[PCI+08] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[Pet04] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2004.

[PKA+07] G. Pavlidis, A. Koutsoudis, F. Arnaoutoglou, V. Tsioukas, and C. Chamzas. Methods for 3D Digitization of Cultural Heritage. *Journal of Cultural Heritage*, 8(1):93–98, 2007.

[PL08] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Information Retrieval*, 2(1-2):1–135, 2008.

[Pla99] J.C. Platt. Probabilities for SV Machines. *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 61–74, 1999.

[Plu80] R. Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, Publishers, 1980.

[PLV02] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs Up?: Sentiment Classification using Machine Learning Techniques. In *Proc. ACL Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[PUB09] R. Paredes, A. Ulges, and T. Breuel. Fast Discriminative Linear Models for Scalable Video Tagging. In *Proc. Int. Conf. on Machine Learning and Applications*, pages 571–576, December 2009.

[QHR+07] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative Multi-Label Video Annotation. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 17–26, September 2007.

[QMO+07] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars. A Thousand Words in a Scene. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007.

[QSH+06] G. Qi, Y. Song, X.S. Hua, H.J. Zhang, and L.R. Dai. Video Annotation by Active Learning and Cluster Tuning. In *Computer Vision and Pattern Recognition (CVPR) Workshop*, pages 114–114, 2006.

[Qui86] J. Ross Quinlan. Induction of Decision Trees. *Machine learning*, 1(1):81–106, 1986.

[RDM08] K. Radinsky, S. Davidovich, and S. Markovitch. Predicting the News of Tomorrow using Patterns in Web Search Queries. In *Proc. Int. Conf. on Web Intelligence and Intelligent Agent Technology (WIC)*, volume 1, pages 363–367, 2008.

[RFF10] O. Russakovsky and L. Fei-Fei. Attribute Learning in Largescale Datasets. In *Proc. ECCV Workshop on Parts and Attributes*, volume 1, September 2010.

[RFF12] O. Russakovsky and L. Fei-Fei. Attribute Learning in Large-scale Datasets. In *Trends and Topics in Computer Vision*, pages 1–14. Springer, 2012.

[RFM10] J. Ratkiewicz, A. Flammini, and F. Menczer. Traffic in Social Media: Paths through Information Networks. In *Proc. IEEE Int. Conf. on Social Computing (SocialCom)*, pages 452–458, September 2010.

[RHC99] Y. Rui, T. Huang, and S.-F. Chang. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, 1999.

[RHW86] D. Rumelhart, G. Hintont, and R. Williams. Learning Representations by Back-propagating Errors. *NATURE*, 323:9, 1986.

[RLFF13] V. Ramanathan, P. Liang, and L. Fei-Fei. Video Event Understanding using Natural Language Descriptions. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 905–912, 2013.

[RM01] N. Roy and A. McCallum. Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 441–448, June 2001.

[RMJ+09] C. Ramachandran, R. Malik, X. Jin, J. Gao, K. Nahrstedt, and J. Han. VideoMule: A Consensus Learning Approach to Multi-label Classification from Noisy User-Generated Videos. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 721–724, October 2009.

[RMZL12] S. Roy, T. Mei, W. Zeng, and S. Li. SocialTransfer: Cross-domain Transfer Learning from Social Streams for Media Applications. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 649–658, October 2012.

[Roc71] J. Rocchio. Relevance Feedback in Information Retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323, 1971.

[Rot08] P. Roth. Survey of Appearance-based Methods for Object Recognition. Technical Report ICG-TR-01/08, Computer Graphics & Vision, TU Graz, 2008.

[RRR+08] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. Torr. Randomized trees for human pose detection. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[RSD+12a] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and Predicting Behavioral Dynamics on the Web. In *Proc. ACM Int. Conf. World Wide Web (WWW)*, pages 599–608, April 2012.

[RSD12b] M. Reif, F. Shafait, and A. Dengel. Meta-Learning for Evolutionary Parameter Optimization of Classifiers. *Machine learning*, 87(3):357–380, 2012.

[RSG+12] M. Reif, F. Shafait, M. Goldstein, T. Breuel, and A. Dengel. Automatic Classifier Selection for Non-Experts. *Pattern Analysis and Applications*, pages 1–14, 2012.

[SB90] G. Salton and C. Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[SC97] John R Smith and Shih-Fu Chang. VisualSEEk: a fully Automated Content-based Image Query System. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 87–98, October 1997.

[SC00] G. Schohn and D. Cohn. Less is More: Active Learning with Support Vector Machines. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 839–846, 2000.

[SCZ07] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting Image Databases from the Web. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 1–8, October 2007.

[SDLW10] N. Sawant, R. Datta, J. Li, and J. Z. Wang. Quest for Relevant Tags using Local Interaction Networks and Visual Content. In *Proc. ACM Int. Conf. on Multimedia Information Retrieval (ACM MIR)*, March 2010.

[Set09] B. Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[SGK+07] F. Seinstra, J. Geusebroek, D. Koelma, C. Snoek, M. Worring, and A. Smeulders. High-performance Distributed Video Content Analysis with Parallel-horus. *IEEE Multimedia*, 14(4):64–75, 2007.

[SH10] G. Szabo and B. Huberman. Predicting the Popularity of Online Content. *Communications of the ACM*, 53(8):80–88, 2010.

[SHBD14] C. Schulze, D. Henter, D. Borth, and A. Dengel. Automatic Detection of CSA Media by Multi-modal Feature Fusion for Law Enforcement Support. In *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR)*, pages 353–361, April 2014.

[SHDW05] Y. Song, X.S. Hua, L.R. Dai, and M. Wang. Semi-automatic Video Annotation based on Active Learning with Multiple Complementary Predictors. In *Proc. ACM Int. Workshop on Multimedia Information Retrieval (MIR)*, pages 97–104, October 2005.

[Sil] D. Silversmith. Google Losing up to $ 1.65M a Day on YouTube. available from internetevolution.com (retrieved: December 2011).

[SJC08] J. Shotton, M. Johnson, and R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[SLNM11] R. Socher, C. C Lin, A. Ng, and C. Manning. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 129–136, June 2011.

[SM86] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[SMB00] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of Interest Point Detectors. *Int. Journal Computer Vision*, 37(2):151–172, 2000.

[Sme05] A. Smeaton. Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pages 11–17, 2005.

[Sme07] A. Smeaton. Techniques Used and Open Challenges to the Analysis, Indexing and Retrieval of Digital Video. *Information Systems*, 32(4):545–559, 2007.

[SMF+12] S. Strassel, A. Morris, J. G Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating HAVIC: Heterogeneous Audio Visual Internet Collection. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 2573–2577, 2012.

[SMH04] F. Souvannavong, B. Merialdo, and B. Huet. Latent Semantic Indexing for Semantic Content Detection of Video Shots. In *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, pages 1783–1786, June 2004.

[SOD10] A. Smeaton, P. Over, and A. Doherty. Video Shot Boundary Detection: Seven Years of TRECVid Activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.

[SOJN08] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it Good?: Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proc. ACL Conf. on Empirical Methods in Natural Language Processing*, pages 254–263, September 2008.

[SOK04] A. Smeaton, P. Over, and W. Kraaij. TRECVID: Evaluating the Effectiveness of Information Retrieval Tasks on Digital Video. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 652–655, October 2004.

[SOK06] A. Smeaton, P. Over, and W. Kraaij. Evaluation Campaigns and TRECVid. In *Proc. ACM Workshop on Multimedia Information Retrieval (MIR)*, pages 321–330, October 2006.

[SOK09] A. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer, 2009.

[SOU09] Netherlands Institute for Sound and Vision. available from `http://instituut.beeldengeluid.nl/` (retrieved: Feb'09), Feburary 2009.

[SPPC+12] S. Samangooei, D. Preotiuc-Pietro, T. Cohn, M. Niranjan, and N. Gibbins. Trendminer: An Architecture for Real Time Analysis of Social Media Text. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, June 2012.

[SREZ05] J. Sivic, B. Russell, A. Efros, and A. Zisserman. Discovering Objects and their Location in Images. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 370–377, October 2005.

[SS01] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[SS02] P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., 2002.

[SS09] A. Setz and C. Snoek. Can Social tagged Images aid Concept-based Video Search? In *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, pages 1460–1463, June 2009.

[SS10] C. Snoek and A. Smeulders. Visual-Concept Search Solved? *IEEE Computer*, 43(6):76–78, June 2010.

[SSD11] A. Shahab, F. Shafait, and A. Dengel. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, pages 1491–1496. IEEE, 2011.

[SSTK08] Y. Sun, S. Shimada, Y. Taniguchi, and A. Kojima. A Novel Region-based Approach to Visual Concept Modeling using Web Images. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 635–638, October 2008.

[SSW07] S. Sengamedu, N. Sawant, and S. Wadhwa. vADeo: Video Advertising System. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 455–456, October 2007.

[SvdSF+13] C. Snoek, K. van de Sande, D. Fontijne, A. Habibian, M. Jain, S. Kordumova, Z. Li, M. Mazloom, S.-L. Pintea, R. Tao, D. Koelma, and A. Smeulders. MediaMill at TRECVID 2013: Searching concepts, objects, instances and events in video. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2013.

[SvZ08] B. Sigurbjörnsson and R. van Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *Proc. ACM Int. Conf. on World Wide Web (WWW)*, pages 327–336, April 2008.

[SW05] C. Snoek and M. Worring. Multimodal Video Indexing: A Review of the State-of-the-Art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[SW09] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Inf. Retrieval*, 4(2), 2009.

[SWdR+08] C. Snoek, M. Worring, O. de Rooij, K. van de Sande, R. Yan, and A. Hauptmann. VideOlympics: Real-Time Evaluation of Multimedia Retrieval Systems. *IEEE MultiMedia*, 15(1):86–91, 2008.

[SWG+05] C. Snoek, M. Worring, J.-M. Geusebroek, D. Koelma, and F. Seinstra. On the Surplus Value of Semantic Video Analysis Beyond the Key Frame. In *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, July 2005.

[SWH+06] C. Snoek, M. Worring, B. Huurnink, J. van Gemert, K. van de Sande, D. Koelma, and O. de Rooij. MediaMill: Video Search using a Thesaurus of 500 Machine Learned Concepts. In *Proc. Int. Conf. Semantic and Digital Media Techn. (SAMT)*, December 2006.

[SWS05] C. Snoek, M. Worring, and A. Smeulders. Early versus Late Fusion in Semantic Video Analysis. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 399–402, Novermber 2005.

[SWSJ00] A. Smeulders, M. Worring, S. Santini, and A. Gupta R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[SWvG+06a] C. Snoek, M. Worring, J. van Gemert, J. Geusebroek, and A. Smeulders. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *Proc. Int. Conf. on Multimedia*, pages 225–226, October 2006.

[SWvG+06b] C. Snoek, M. Worring, J. van Gemert, J. Geusebroek, and A. Smeulders. The MediaMill TRECVID 2006 Semantic Video Search Engine. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2006.

[SWWdR08] A. Smeaton, P. Wilkins, M. Worring, and O. de Rooij. Content-Based Video Retrieval: Three Example Systems from TRECVid. *Int. Journal of Imaging Science and Technology*, 18(2-3):195–201, 2008.

[SZ03] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 1470–1477, October 2003.

[SZ06] J. Sivic and A. Zisserman. Video Google: Efficient Visual Search of Videos. In *Toward Category-Level Object Recognition*, pages 127–144. Springer, 2006.

[TAP+10] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on YouTube: Tag Recommendation and Category Discovery. In *Int. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2010.

[TBP+10] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[TC01] S. Tong and E. Chang. Support Vector Machine Active Learning for Image Retrieval. In *Proc. ACM Int. Conf. on Multimedia (ACM)*, pages 107–118, September 2001.

[TFF08] A. Torralba, R. Fergus, and W. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[TK91] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical report, Carnegie Mellon University, 1991.

[TK02] S. Tong and D. Koller. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research (JMLR)*, 2:45–66, 2002.

[TMY78] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Trans. System, Man, Cybernetics*, 8(6):460–472, 1978.

[TRE07] TRECVID-2007: Shot Boundary Detection Task Summary. Proc. NIST TRECVID Workshop, November 2007.

[TSF10] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient Object Category Recognition using Classemes. In *Proc. European. Conf. on Computer Vision (ECCV)*, pages 776–789. Springer, 2010.

[TSSW10] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, pages 178–185, May 2010.

[Tur02] P. Turney. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proc. ACL Annual Meeting on Association for Computational Linguistics*, pages 417–424, 2002.

[TVLK12] M. ten Thij, Y. Volkovich, D. Laniado, and A. Kaltenbrunner. Modeling and Predicting Page-view Dynamics on Wikipedia. *arXiv preprint arXiv:1212.5943*, 2012.

[TYH+09] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring Semantic Concepts from Community-contributed Images and Noisy Tags. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 223–232, October 2009.

[UBB10] A. Ulges, D. Borth, and T. Breuel. Visual Concept Learning from Weakly Labeled Web Videos. In *Video Search and Mining*, pages 203–232. Springer, 2010.

[UBK13] A. Ulges, D. Borth, and M. Koch. Content Analysis Meets Viewers: Linking Concept Detection with Demographics on YouTube. *Int. Journal of Multimedia Information Retrieval*, pages 1–13, 2013.

[UKB12] A. Ulges, M. Koch, and D. Borth. Linking Visual Concept Detection with Viewer Demographics. In *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR)*, pages 24–32, June 2012.

[UKBB09] A. Ulges, M. Koch, D. Borth, and T. Breuel. TubeTagger - YouTube-based Concept Detection. In *Proc. Int. Workshop on Internet Multimedia Mining*, pages 190–195, December 2009.

[UKSB08] A. Ulges, M. Koch, C. Schulze, and T. Breuel. Learning TRECVID'08 High-level Features from YouTube. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, November 2008.

[Ulg09] A. Ulges. *Visual Concept Learning from User-tagged Web Video.* PhD thesis, University of Kaiserslautern, Germany, 2009.

[US11] A. Ulges and A. Stahl. Automatic Detection of Child Pornography using Color Visual Words. In *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2011.

[USBS12] A. Ulges, C. Schulz, D. Borth, and A. Stahl. Pornography Detection in Video Benefits (a lot) from a Multi-modal Approach. In *Proc. ACM Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis (AMVA)*, pages 21–26, October 2012.

[USKB08a] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. A System that Learns to Tag Videos by Watching Youtube. In *Proc. Int. Conf. on Vision Systems*, pages 415–424, May 2008.

[USKB08b] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. Identifying Relevant Frames in Weakly Labeled Videos for Training Concept Detectors. In *Proc. ACM Int. Conf. Image and Video Retrieval (CIVR)*, pages 9–16, July 2008.

[USKB10] A. Ulges, C. Schulze, M. Koch, and T. Breuel. Learning Automatic Concept Detectors from Online Video. *Computer Vision Image Understanding*, 114(4):429–438, 2010.

[VA06] L. Von Ahn. Games with a Purpose. *IEEE Computer*, 39(6):92–94, 2006.

[VABHL03] L. Von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using Hard AI Problems for Security. In *Advances in Cryptology (EUROCRYPT)*, pages 294–311. Springer, Mai 2003.

[VAD08] L. Von Ahn and L. Dabbish. Designing Games with a Purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[Vap00] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 2000.

[VD02] R. Vilalta and Y. Drissi. A Perspective View and Survey of Meta-Learning. *Artificial Intelligence Review*, 18(2):77–95, 2002.

[vdSGS08a] K. van de Sande, T. Gevers, and C. Snoek. A Comparison of Color Features for Visual Concept Classification. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pages 141–150, July 2008.

[vdSGS08b] K. van de Sande, T. Gevers, and C. Snoek. Evaluation of Color Descriptors for Object and Scene Recognition. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[vdSGS10] K. van de Sande, T. Gevers, and C. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[vGV+06] J. van Gemert, J. Geusebroek, C. Veenman, C. Snoek, and A. Smeulders. Robust Scene Categorization by Learning Image Statistics in Context. In *CVPR Workshop on Semantic Learning Applications in Multimedia*, June 2006.

[vGVSG10] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual Word Ambiguity. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[VH+05] E. Voorhees, D. K Harman, et al. *TREC: Experiment and Evaluation in Information Retrieval*, volume 63. MIT Press Cambridge, 2005.

[VW12] V. Vonikakis and S. Winkler. Emotion-based Sequence of Family Photos. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1371–1372, October 2012.

[WA12] S. Whiting and O. Alonso. Hashtags as Milestones in Time. In *Proc. Int. Workshop on Time-aware Information Access (TAIA)*, 2012.

[WBB11] K. Wang, B. Babenko, and S. Belongie. End-to-End Scene Text Recognition. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 1457–1464, 2011.

[WCGH99] H. Wactlar, M. Christel, Y. Gong, and A. Hauptmann. Lessons Learned from Building a Terabyte Digital Video Library. *IEEE Computer*, 32(2):66–73, 1999.

[WCLH11] O. Wu, Y. Chen, B. Li, and W. Hu. Evaluating the Visual Quality of Web Pages using a Computational Aesthetic Approach. In *Proc. ACM Int. Conf. on Web Search and Data Mining (WSDM)*, pages 337–346, February 2011.

[WH08] W. Wang and Q. He. A Survey on Emotional Semantic Image Retrieval. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pages 117–120, October 2008.

[WHS+06] M. Wang, X.-S. Hua, Y. Song, X. Yuan, S. Li, and H.-J. Zhang. Automatic Video Annotation by Semi-supervised Learning with Kernel Density Estimation. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 967–976, October 2006.

[WJH+12] X. Wang, J. Jia, P. Hu, S. Wu, J. Tang, and L. Cai. Understanding the Emotional Impact of Images. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 1369–1370, October 2012.

[WJN09] X.-Y. Wei, Y.-G. Jiang, and C.-W. Ngo. Exploring Inter-Concept Relationship with Context Space for Semantic Video Indexing. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, page 15, July 2009.

[WJR06] R. White, J. Jose, and I. Ruthven. An Implicit Feedback Approach for Interactive Information Retrieval. *Information Processing and Management*, 42(1):166–190, 2006.

[WL11] J. Weng and B.-S. Lee. Event Detection in Twitter. In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, pages 401–408, July 2011.

[WLL+07] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video Diver: Generic Video Indexing with Diverse Features. In *Proc. Int. Workshop on Multimedia Information Retrieval (MIR)*, pages 61–70, September 2007.

[WLLZ07] D. Wang, X. Li, J. Li, and B. Zhang. The Importance of Query-Concept-Mapping for Automatic Video Retrieval. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 285–288, September 2007.

[WN07] B. Wu and R. Nevatia. Improving Part based Object Detection by Unsupervised, Online Boosting. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[WN08] X.-Y. Wei and C.-W. Ngo. Fusing Semantics, Observability, Reliability and Diversity of Concept Detectors for Video Search. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 81–90, October 2008.

[WR88] J. Wiebe and W. Rapaport. A Computational Theory of Perspective and Reference in Narrative. In *Proc. ACL Annual Meeting on Association for Computational Linguistics*, pages 131–138, 1988.

[WR05] J. Wiebe and E. Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proc. Int. Conf. on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 486–497. Springer, 2005.

[WS08] K. Wnuk and S. Soatto. Filtering Internet Image Search Results Towards Keyword Based Category Recognition. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[WSC+12] Z. Wang, L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang. Propagation-Based Social-Aware Replication for Social Video Contents. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, October 2012.

[WWB01] J. Wiebe, T. Wilson, and M. Bell. Identifying Collocations for Recognizing Opinions. In *Proc. Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, pages 24–31, 2001.

[WWC05] J. Wiebe, T. Wilson, and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[WWH05] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proc. Conf. on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, October 2005.

[WYZ+09] X.-J. Wang, M. Yu, L. Zhang, R. Cai, and W.-Y. Ma. Argo: Intelligent Advertising by Mining a User's Interest from his Photo Collections. In *Proc. KDD Workshop on Data Mining and Audience Intelligence for Advertising*, pages 18–26, 2009.

[XC96] J. Xu and W. Croft. Query Expansion using Local and Global Document Analysis. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (ACM SIGIR)*, pages 4–11, August 1996.

[XW09] C.W. Ngo X. Wu. Towards Google Challenge: Combining Contextual and Social Information for Web Video Categorization. In *Proc. ACM Int. Conf. on Multimedia (ACM MM), Multimedia Grand Challenge*, pages 1109–1110, October 2009.

[YA06] E. Yilmaz and J. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 102–111, November 2006.

[YB05] K. Yanai and K. Barnard. Probabilistic Web Image Gathering. In *Int. Workshop on Multimedia Information Retrieval (MIR)*, pages 57–64, November 2005.

[YCKH07] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Technical report, Columbia University, 2007.

[YH06] R. Yan and A. Hauptmann. Probabilistic Latent Query Analysis for Combining Multiple Retrieval Sources. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (ACM SIGIR)*, pages 324–331, August 2006.

[YH07] R. Yan and A. Hauptmann. A Review of Text and Image Retrieval Approaches for Broadcast News Video. *Information Retrieval*, 10(4):445–484, 2007.

[YH08a] J. Yang and A. Hauptmann. A Framework for Classifier Adaptation and its Applications in Concept Detection. In *Proc. ACM Int. Conf. on Multimedia Information Retrieval (ACM MIR)*, pages 467–474, 2008.

[YH08b] J. Yang and A. Hauptmann. (Un)Reliability of Video Concept Detection. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pages 85–94, 2008.

[YJT+12] F. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak Attributes for Large-scale Image Retrieval. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2949–2956, June 2012.

[YL11] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proc. ACM Int. Conf. Web Search and Data Mining (WSDM)*, pages 177–186, February 2011.

[YMH+07] B. Yang, T. Mei, X. Hua, L. Yang, S. Yang, and M. Li. Online Video Recommendation Based on Multimodal Fusion and Relevance Feedback. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pages 73–80, July 2007.

[YN07] C. Yang and T. Ng. Terrorism and Crime related Weblog Social Network: Link, Content Analysis and Information Visualization. In *Proc. IEEE Intelligence and Security Informatics (ISI)*, pages 55–58, May 2007.

[YNTT11] K. Yatani, M. Novati, A. Trusty, and K. Truong. Analysis of Adjective-Noun Word Pair Extraction Methods for Online Review Summarization. In *Proc. AAAI Int. Conf. on Artificial Intelligence*, pages 2771–2776. AAAI Press, 2011.

[YOU11] Bringing German History Online. available from http://youtube-global.blogspot.de/2011/10/bringing-german-history-online.html (retrieved: Oct'12), October 2011.

[YOU13] YouTube Press Statistics. available from http://www.youtube.com/yt/press/statistics.html (retrieved: Aug'13), August 2013.

[YSR05] A. Yavlinsky, E. Schofield, and S. Rüger. Automated Image Annotation using Global Features and Robust Nonparametric Density Estimation. In *Proc. ACM Int. Conf. on Image and Video Retrieval (CIVR)*, pages 507–517, July 2005.

[YT11] W. Yang and G. Toderici. Discriminative Tag Learning on YouTube Videos with Latent Sub-Tags. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3217–3224. IEEE, 2011.

[YUB+12] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe. In the Eye of the Beholder: Employing Statistical Analysis and Eye Tracking for Analyzing Abstract Paintings. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 349–.358, October 2012.

[YvGR+08] V. Yanulevskaya, J. van Gemert, K. Roth, A. Herbold, N. Sebe, and J.M. Geusebroek. Emotional Valence Categorization using Holistic Image Features. In *Proc. IEEE Int Conf on Image Processing (ICIP)*, pages 101–104, October 2008.

[YWX+07] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A Formal Study of Shot Boundary Detection. *IEEE Trans. Circuits and Systems for Video Technology*, 17(2):168–186, 2007.

[YYH04] R. Yan, J. Yang, and A. Hauptmann. Learning Query-class Dependent Weights in Automatic Video Retrieval. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 548–555, October 2004.

[ZMLS07] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *Int. Journal Computer Vision*, 73(2):213–238, 2007.

[ZRHM98] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra. Adaptive Key Frame Extraction using uUnsupervised lustering. In *IEEE Int. Conf. on Image Processing (ICIP)*, volume 1, pages 866–870. IEEE, 1998.

[ZSC+06] R. Zhang, R. Sarukkai, J.H. Chow, W. Dai, and Z. Zhang. Joint Categorization of Queries and Clips for web-based Video Search. In *Proc. ACM Int. Workshop on Multimedia Information Retrieval (MIR)*, pages 193–202, October 2006.

[ZWIL00] D. Zhang, A. Wong, M. Indrawan, and G. Lu. Content-based Image Retrieval using Gabor Texture Features. In *IEEE Pacific-Rim Conf. on Multimedia*, 2000.

Curriculum Vitae

Name: Damian Borth

Education

School Education: Carl-Benz Schule, Mannheim,
*A*bitur, 2001

University Education: University of Kaiserslautern
Master in Computer Science (University of Kaiserslautern), 2010
University of Corporate Education, Mannheim
Diploma in Engineering (BA) (University of Corporate Education), 2004

Academic and Professional Experience

Oct 2007 – Mar 2014	Fellow of PhD program of computer science at the University of Kaiserslautern,
	PhD Fellowship of the Max Planck Institute of Informatics, Kaiserslautern
Aug 2012 – Dec 2012	Visiting Scholar, Columbia University, New York, USA
Aug 2003 – Dec 2003	Visiting Student, University of California, Santa Barbara, CA, USA