

Technische Universität Kaiserslautern  
Fachbereich Informatik

Masterarbeit

**Ad Targeting for Web Video  
by Automatic Video Annotation**

von

Markus Koch

31. Oktober 2011

Betreuer:  
Prof. Dr. Prof. h.c. Andreas Dengel  
Dr. Adrian Ulges



## **Erklärung**

Ich versichere hiermit, dass ich die vorliegende Masterarbeit mit dem Thema "Ad Targeting for Web Video by Automatic Video Annotation" selbstständig verfasst habe und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, habe ich durch die Angabe der Quelle, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht.

Kaiserslautern, den 31. Oktober 2011

Unterschrift



## **Kurzfassung**

Videoportale wie YouTube oder Vimeo haben in den vergangenen Jahren ein starkes Wachstum verzeichnet. Die hohe Zahl an Nutzern und an verfügbarem Videomaterial setzt nicht nur eine effiziente Suche voraus, sondern erfordert auch immer öfters die Einbindung von Werbung, um die Kosten für Bandbreite und Speicherplatz zu decken. Dadurch, daß viele Videos nicht ausreichend annotiert sind, wird die gezielte Einbindung von Werbung jedoch erheblich erschwert.

Um diesem Problem zu begegnen, untersucht diese Arbeit die Beziehung zwischen dem Vorhandensein semantischer Konzepte (zum Beispiel Objekte oder Aktivitäten) in einem Video sowie dem demographischen Profil der Zuschauer. Es wird gezeigt, daß hier oftmals eine starke Verbindung existiert.

Basierend auf dieser Beobachtung werden zwei Ideen vorgestellt: Erstens, ein System, welches die Ergebnisse einer Konzepterkennung nutzt, um das demographische Profil eines Videos zu schätzen. Diese Information kann dann dazu genutzt werden, um darauf zugeschnittene Werbung zu empfehlen oder um die Annotation des Videos zu ergänzen. Zweitens, wenn auf die Demographie eines Videos bereits zugegriffen werden kann (zum Beispiel über Zuschauerdaten), die Integration dieser mit einem Konzepterkennungssystem, um dessen Genauigkeit zu steigern.

## **Abstract**

Over the last years, web video portals like YouTube or Vimeo have seen an immense growth in user numbers and available video content. Consequently, they not only require an effective video search and recommendation, but also the integration of online advertising to bear the high costs for storage and traffic. This, however, is often difficult to achieve, as many videos come with little to no textual meta-data.

To overcome this problem, this thesis explores the connection between semantic concepts (like objects or activities) appearing in a video and the demographics of the potential viewership. It is shown that semantic concepts are often strongly related to specific demographic groups.

Based on this observation, two strategies are proposed: First, an approach that estimates the demographic distribution of the viewership based on the results of a concept detection, that is by only using visual features obtained from the video. This demographic information can then be used for advertising purposes, or to enrich the video's annotation. And second, if demographic context is available for a video, it is proposed to integrate it as an additional modality in concept detection to improve concept detection accuracy.



# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
<b>2</b>	<b>Related Work</b>	<b>11</b>
2.1	Online Advertising . . . . .	11
2.2	Image and Video Content Analysis for Advertising . . . . .	13
2.3	Demographics Prediction . . . . .	14
2.4	Concept Detection . . . . .	14
<b>3</b>	<b>Video Concept Detection</b>	<b>17</b>
3.1	Overview . . . . .	17
3.2	Feature Extraction . . . . .	18
3.2.1	Color Correlograms . . . . .	18
3.2.2	Bag of Visual Words . . . . .	19
3.3	Classifiers . . . . .	20
3.3.1	Support Vector Machines . . . . .	20
3.3.2	PAMIR . . . . .	24
<b>4</b>	<b>Inferring Demographic Profiles by Clustering</b>	<b>27</b>
4.1	Idea . . . . .	27
4.2	Demographic Profiles . . . . .	29
4.3	K-means . . . . .	31
4.4	Clustering Results . . . . .	33
<b>5</b>	<b>Linking Video Content with Demographics</b>	<b>37</b>
5.1	Overview . . . . .	37
5.2	Demographics Estimation by Concept Detection . . . . .	37
5.2.1	Baseline System . . . . .	38
5.2.2	Marginalization-based Approach . . . . .	39
5.2.3	The Two-step Approach . . . . .	41
5.3	Demographic Profiles for Improving Concept Detection . . . . .	43
5.3.1	Early Fusion . . . . .	43
5.3.2	Late Fusion . . . . .	44

<b>6 Experiments</b>	<b>47</b>
6.1 Dataset . . . . .	47
6.2 Experiments . . . . .	49
6.2.1 Concept detection . . . . .	49
6.2.2 Demographics estimation . . . . .	54
<b>7 Discussion</b>	<b>63</b>
<b>Bibliography</b>	<b>67</b>
<b>Appendix</b>	<b>A-1</b>



# List of Figures

1.1	Example videos without meaningful keywords, title or description.	8
2.1	Example implementations of two advertising strategies. . . . .	12
2.2	TubeTagger retrieval demonstration. . . . .	15
3.1	Concept detection overview. . . . .	18
3.2	Bag of visual words. . . . .	20
3.3	Support Vector Machine example. . . . .	21
4.1	Example for demographic targeting. . . . .	28
4.2	Video views versus comment count. . . . .	29
4.3	Obtaining a demographic profile. . . . .	31
4.4	K-means example. . . . .	32
4.6	Visual impressions for seven demographic clusters. . . . .	36
5.1	Overview of the baseline system. . . . .	39
5.2	Overview of the marginalization-based system. . . . .	41
5.3	Overview of the two-step approach. . . . .	42
6.1	Concept detection results comparing a random split into train and test data with a split by the number of comments. . . . .	50
6.2	Top rated keyframes for two concepts where the use of demographic context helps improving concept detection results. . . . .	55
6.3	Demographics estimation results for the baseline system. . . . .	56
6.4	Videos from two different demographic clusters. . . . .	57
6.5	Demographics estimation results for the marginalization system. . . . .	58
6.6	Two-step demographics estimation results. . . . .	59
6.7	Change of mean average precision for the demographics estimate when removing 'bad' concepts. . . . .	61
7.1	Examples that show some of the problems with the dataset. . . . .	64



# List of Tables

4.1	Demographic clustering results. . . . .	35
6.1	Age and gender distribution in the YouTube dataset. . . . .	48
6.2	Concept detection results for different combinations of classifiers and features. . . . .	51
6.3	Concept detection results using demographic profiles as a feature. . . . .	52
6.4	Concept detection results that show the highest and lowest difference in average precision, comparing visual words and demographic profiles. . . . .	53
6.5	Concept detection results for the fusion of visual and demographic features. . . . .	54
A.1	Listing of semantic concepts. . . . .	A-1
A.2	Detailed results for the concept detection experiments. . . . .	A-7
A.3	Detailed results for the feature fusion experiments. . . . .	A-10
A.4	Detailed results for the baseline system. . . . .	A-13
A.5	Detailed results for the marginalization-based approach. . . . .	A-13
A.6	Detailed results for the two-step system. . . . .	A-14



# Chapter 1

## Introduction

### Motivation

For several years now, video portals like YouTube<sup>1</sup>, Vimeo<sup>2</sup> or MyVideo<sup>3</sup> have experienced an immense growth in user numbers. More people than ever before are using these platforms to share private moments, to spread information, or just for personal entertainment. Especially amongst younger generations, web video is gradually replacing television as the standard video broadcast medium, as it offers an easy way to share and access digital videos world-wide.

In 2007, the market leader YouTube already hosted about 72 million videos with approximately 200,000 new videos a day. During the past 4 years, this number has increased to a staggering 295 million videos [1]. Today, more than 48 hours of video material are uploaded to YouTube every minute [2]. But while this number is impressive, it leads to several major problems, both for the consumer and the owner of the web portal.

One problem is that maintaining a large video database is very expensive. Storing video material requires huge amounts of hard disk capacity, and serving millions of views every day (YouTube: 700 billion views in the year 2010 [66]) also causes high bandwidth and traffic costs. The introduction of high-definition video up to a resolution of 1080p and the rapid spread of broadband internet access have further increased this.

To address this issue, a video portal can be monetized by integrating online advertising with both the web site and the video player. As the key economic driver of the internet, online advertising already founded uncountable web services and communities. In the first half of 2011, internet advertising revenues reached

---

<sup>1</sup><http://www.youtube.com>

<sup>2</sup><http://www.vimeo.com>

<sup>3</sup><http://www.myvideo.de>



(a) Horse riding

(b) Soccer

Figure 1.1: Example videos without meaningful keywords, title or description. The lack of meta-data makes it hard to retrieve these videos, to make recommendations, or to select appropriate advertisements.

a record volume of nearly \$15 billion, with the rate of growth doubling year-over-year [19]. Therefore, online advertising not only has the potential to fund large video portals, but can also turn them profitable.

A successful advertising strategy requires the user to be interested in the products and services advertised, so the method of how the advertisements are tailored to the customer has a large impact on their success. To do so, most online advertising systems still rely on textual information, like the user’s query in case of a web search or keywords extracted from the web site the user is currently browsing. This approach works well with generic web sites, but it is limited when applied to a web video portal, as the textual information available is often unreliable. Videos might be annotated insufficiently or completely lacking any textual description (see Figure 1.1).

In addition, target groups in advertising are often defined by demographic attributes like age and gender, and this information is often not available about the audience of a web video. In particular, freshly uploaded videos are lacking any representative view data or other means to infer demographic interest.

## Idea

To address this issue, a system is proposed that analyses the visual content of a video and uses this information to estimate the demographics of the video’s audience. The visual analysis is performed by a *concept detection* system, which detects the presence of semantic concepts like objects or activities based on a

low-level visual description of the video.

Knowing about the demographics of potential viewers, one can either directly recommend appropriate advertisements or employ this information as an additional input for an existing advertising system. In addition, the concept detection can suggest a category for the video, which can be used for a further customization of the advertising process. Since this approach only uses low-level visual features, it can be applied even if the video was just recently uploaded, thus no other means of obtaining this information are available.

Furthermore, the results of the visual content analysis can be used to address other common challenges to web video portals. Because browsing and searching video databases still relies on user-generated annotations, which is a task often neglected by the uploader of a video, many videos are difficult to retrieve. For example, only 30 % of all videos on YouTube account for 99 % of the views [66], which is partially because of a sparse annotation. While this is currently being addressed by researchers (e.g. TagSuggest [3]), incorporating available demographic information could further improve the results. Therefore, this thesis also proposes the use of a demographic feature for video concept detection, and evaluates its performance.

In summary, this thesis proposes the following approaches to address the previously mentioned problems:

- A system that predicts the demographics of a video’s audience by using only the visual content of the video. This is achieved by the means of a concept detection system, together with an initially trained demographic model, using demographic data available through user comments on the web video portal YouTube.
- A concept-detection system that uses demographic context available through user comments as an additional modality in concept detection (alongside traditional feature representations like color and visual words).

These systems are evaluated in a series of experiments, using more than 14,000 videos automatically downloaded from the video portal YouTube, and the user profiles of 2,2 million users that commented on these videos. The experimental results lead to the following observations:

- There is a strong connection between semantic concepts and the demographics of the audience. For example, the concept ‘skateboarding’ is predominantly viewed by a young male audience, while the concept ‘cake’ attracts mostly female viewers.
- Using concept detection is a suitable approach to estimate the demographics of a videos’ audience.

- Concept detection accuracy can be improved significantly by using demographic context as an additional modality in the classification process. For a state-of-the-art concept detection system, a relative improvement of nearly 50 % could be achieved.

The remainder of this thesis is organized as follows: In Section 2, the research areas related to this thesis are introduced. In Section 3, the technical details of the classifiers and visual features used are explained. Sections 4 and 5 discuss how to obtain and describe the demographic information mentioned above, and explain the proposed approaches in more detail. The experimental results together with a description of the data used can be found in Section 6, followed by a discussion of the results and possible future research directions in Section 7.



# Chapter 2

## Related Work

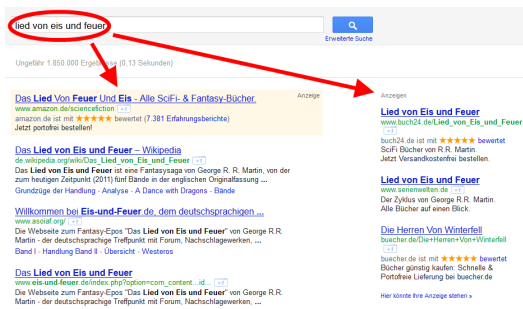
### 2.1 Online Advertising

Compared to other media like print or television, online advertising offers the possibility to target large audiences at reasonable costs. As an indispensable tool for the internet's economic system, it became one of the most important advertising strategies. Today, online advertising founds web communities, free email and streaming services, blogs, and many other web sites. Given this huge influence in the world of modern communication systems, online advertising is an area of ongoing research, with a lot of effort done towards a more effective monetization on the internet.

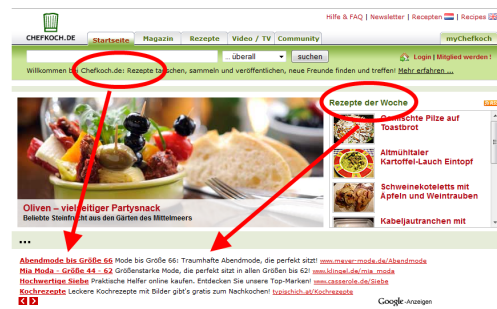
The research in online advertising can be classified into three areas:

- **Keyword-targeted advertising**, where the advertisements are selected based on keywords supplied by the user (for example, by performing a web search).
- **Contextual advertising**, where the selection is based on the content of a web page.
- **Behavioural advertising**, which relies on user profiles and demographics.

The most commonly used approach is the keyword-targeted advertising or 'Sponsored Search'. In this scenario, the keywords within a user's query are used to find and select an appropriate advertisement. This method is usually used within the context of a web search engine where - after the query has been evaluated - the selected advertisements are presented together with the search results (see Figure 2.1). Because there is only a weak connection to the approach presented in this thesis, it is referred to the work of Jansen and Mullen [33] for more information about recent research in this area.



(a) Keyword-targeted advertising



(b) Content-targeted advertising

Figure 2.1: Example implementations of two advertising strategies. (a) Keyword-targeted advertising: the advertisements are tailored according to the performed web search (b) Content-targeted advertising: the advertisements are tailored according to text extracted from the web site.

Keyword-targeted advertising can only be used if the user supplies the necessary textual input. If this option is not available, for example, when a user is just browsing a web site, the advertisements have to be selected in a different way. This led to the development of other kinds of advertising systems, namely behavioural advertising and contextual advertising.

In behavioural advertising, the individual user's web search and browsing behaviours are observed to decide on a selection of advertisements. This is usually done by tracking a user through his user profile or secondary sources like server logs and web cookies. Again, there is only a weak connection to this thesis, as the presented approach uses the visual content of a video and does not maintain a user model. However, many behavioural approaches show the advantages of an automatic segmentation of the user base by clustering [60, 62], which will also be applied in this thesis to generate a demographics model. Here, instead of clustering a set of users, it is used to identify videos that have a similar audience in terms of demographics (see Chapter 4).

The idea behind contextual advertising is to decide on a selection of advertisements based on the contents of a web page, usually the text that can be found within that page. Typical systems first search the target page for prominent keywords, which are then matched against a database of advertisements. The finally selected advertisements are then displayed within the content of the page. A prominent application using this method is Google AdSense<sup>1</sup>, see Figure 2.1 b.

Most of the research in this area focuses on the extraction of appropriate keywords from a web site, like the systems proposed in [47, 61, 65], and on improving

<sup>1</sup><http://google.de/adsense>

the following selection process [11, 37]. Only recently, researchers have begun to utilize other content types that can be found on web sites, like images and videos. The use of content in these systems shows some similarities to the approach presented in this thesis, so they will be explained in more detail.

## 2.2 Image and Video Content Analysis for Advertising

Traditional contextual advertising aims at analysing the text on a web site to suggest advertisements. However, text-based methods can handle only a very small portion of web photos and videos, depending on the available annotations. On image and video portals a large percentage of the content exists in non-textual form, so only a small part of the available information is eventually considered in the advertising process. Therefore, a new research direction aims at using these new modalities to expand on the original idea of contextual advertising.

A first image-based approach was proposed by Mei et al. [40]. Their system, ImageSense, first segments a web page into textual and visual content. The visual content is then analysed using a concept-detection system, and the recognized concepts are used together with the text on the web page to select a set of candidate advertisements. These image advertisements are then compared to the original image to select one that is visually similar, and to find a suitable, unimportant position to place it at.

A similar system, VideoSense [41], was proposed to work on video content. In addition to using textual and visual features like ImageSense, this system also utilizes the audio stream available in most videos to search for an appropriate advertisement. They also state that video ads should not be placed at the beginning or the end of a video, but instead at less intrusive positions. Therefore, they identify insertion points within a video such that an advertisement can be integrated seamlessly, at the most appropriate position.

vADeo [51] also aims at providing contextual advertising for videos. In contrast to placing the advertisement at an arbitrary position within video, they find scene-changes that are suitable for ad placement. In addition, they propose a bookmarking mechanism to avoid disturbing the users' viewing experience.

While all these systems incorporate visual information in the advertisement selection process, they still heavily rely on textual descriptions. The approach proposed in this thesis tries to avoid this by using the visual information to estimate the demographics of a videos' audience, therefore it can also be applied when the video is not surrounded by meaningful text.

## 2.3 Demographics Prediction

Demographics play an important role in the selection and presentation of advertisements. This information could easily be obtained from user profiles on a web site, but many internet users avoid publicly sharing private information like their age or gender. This leads to the idea of estimating a users' demographics from other available data.

A lot work in the area this focused on finding correlations between writing styles and the age or gender of the author. Koppel et al. [49] found out that there exist significant differences between blog authors, so that an automatic classification into different demographic groups is possible. A similar approach could also be applied to the comment data used within this thesis, but instead the focus lies on the visual content analysis.

Other researchers focused on the users' behaviour to generate a prediction. A system proposed by Baglioni et al. [5] predicts the gender and interests of a user by analysing the server logs on a web portal. While their gender prediction only shows a slight improvement over random choice, their strategy for estimating the user's interest could achieve good precision values. Hu et al. [28] proposed the use of web page view information to predict age and gender of an user, and could achieve solid results using their method.

This thesis proposes the use of a visual content analysis to estimate the demographics of a video's audience. To the best of our knowledge, no previous work in this area exists.

## 2.4 Concept Detection

Concept detection is one of the core components of this thesis, therefore some important developments in this field of research will be mentioned in the following. A more detailed overview of the recent research in this area can be found in [54]. The specific machine learning techniques and features applied in this thesis are explained in Section 3.

Visual concept detection is an area of research that has seen a lot of attention over the past years, as the need for an effective indexing and retrieval of videos is growing with the size of the video databases. The basic idea of a concept detection system is to use low-level visual features like color and texture to represent videos, and then associate them with semantic concepts like 'soccer', 'dogs' or 'poker'. This allows users to search videos by text queries, just like regular documents, or supports them with annotation.

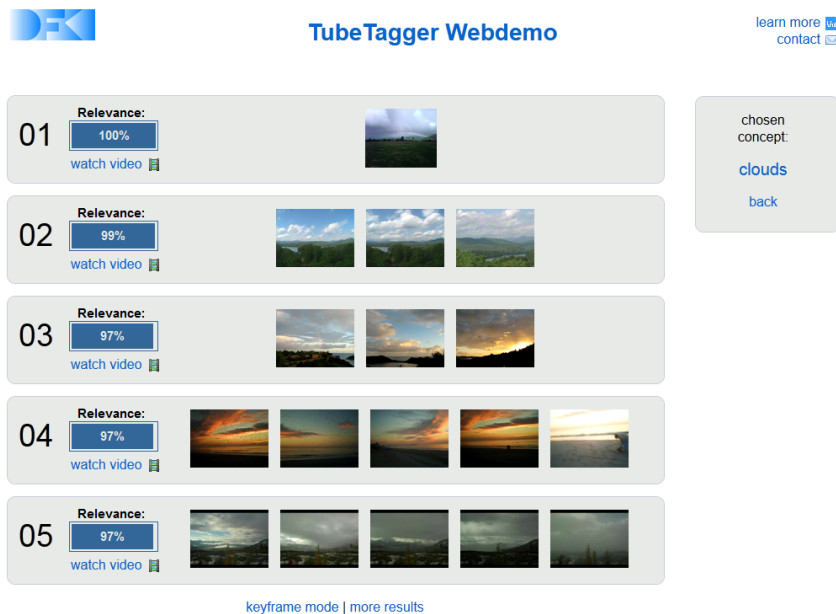


Figure 2.2: Retrieval result for the query 'clouds' on TubeTagger [57]. The videos are ranked according to their relevance to the query.

Many projects have emerged in the context of TRECVID<sup>2</sup>, an annual video retrieval contest that aims at evaluating and promoting the state-of-the-art in this field of research. In 2006, Snoek et al. have run an experiment using 101 semantic concepts [55]. One year later, the University of Columbia introduced a similar system using 374 concepts [63]. Both systems were trained on the dataset provided by TRECVID.

With TubeTagger, a similar system has been developed by Ulges et al. [57] (see Figure 2.2). In contrast to the systems mentioned before, TubeTagger uses a training set automatically obtained from the web video portal YouTube, using the tags and categories provided by the users for annotation. By this, a costly manual annotation can be avoided. This concept detection system uses a vocabulary of 234 concepts, which forms the basis semantic vocabulary for the approach presented in this thesis.

While many concept detection systems use visual features, some researchers evaluated other feature types like audio features [6, 59] or textual features [18, 46] and their use in multi-modal configurations. For example, Yang et al. [64] proposed the simultaneous use of audio, textual and low-level visual features to detect a total of 17 concepts on a dataset obtained from YouTube. So far, a demographic feature similar to the one proposed in this thesis has not been applied to concept detection.

<sup>2</sup><http://trecvid.nist.gov/>

Most concept detection systems, including the one used in this thesis, have a fixed semantic vocabulary to be learned (for example, using parts of the LSCOM ontology<sup>3</sup>, like [63]). However, such predefined concept sets are not optimal because their construction is time-consuming and they do not scale well with the increasing diversity of multimedia content. A solution was proposed by Aradhye [4], which automatically discovers a vocabulary on a set of 25 million videos from YouTube. Although this approach would help in creating a more suitable vocabulary for the proposed demographics estimation, it is beyond the scope of this thesis. Therefore, the experiments will focus on an evaluation using a fixed set of semantic concepts.

---

<sup>3</sup><http://www.lsc.com/ontology/index.html>

# Chapter 3

## Video Concept Detection

### 3.1 Overview

The idea of concept detection is to decide on the presence of a semantic concept (like objects or activities) within an image or video: for example, how likely it is that a dog can be seen in a given image. This is usually treated as a machine learning problem: based on a mathematical description of the images or videos, the so-called features, a set of classifiers is learned over a predefined vocabulary of concepts. In the case of videos, a set of representative images (referred to as keyframes from now) has to be extracted first.

The basic scheme can be seen in Figure 3.1. There are two phases in concept detection: a training phase and a testing phase. In the training phase, the feature vectors for a set of labeled training samples are extracted, which are then used to learn a supervised classifier for each semantic concept in the vocabulary. In the testing phase, we can now infer the presence of a concept  $c$ . For an unseen sample  $X$  we first extract its feature representation  $x$  and apply the previously learned classifier. This results in a score  $P(c|x)$ , describing the probability that concept  $c$  is present in  $X$ .

There exist many different feature types and classifiers, so the number of possible combinations is very high. In this thesis, the focus lies on two visual features that have proven successful in image- and video concept detection: color correlograms and bag-of-visual-word features. Support Vector Machines and PAMIR are used as classifiers. Support Vector Machines mark the state-of-the-art in machine learning, achieving high accuracies in many different tasks, while PAMIR is a very fast model, which is evaluated as an alternative because of scalability reasons.

All features and classifiers used in the context of this thesis are explained in the following sections.

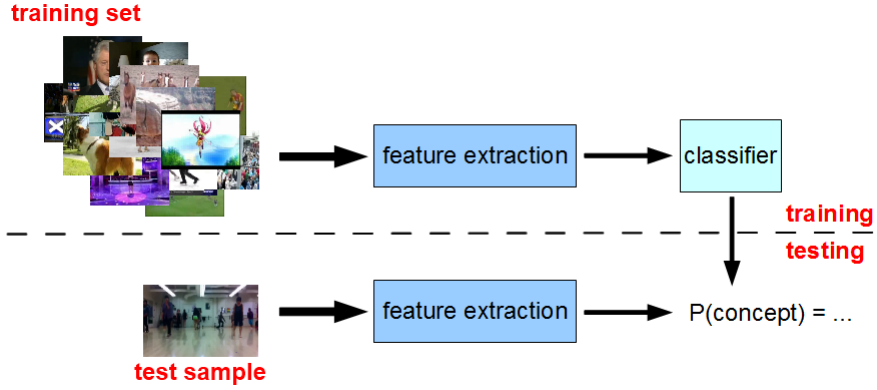


Figure 3.1: General concept detection scheme. Features are extracted from a training set (here: video keyframes) and used to learn a classifier. A new image is then tested by extracting its features and applying the previously learned classifier.

## 3.2 Feature Extraction

### 3.2.1 Color Correlograms

A simple color-based feature is the color histogram, which describes the global color distribution in an image. While it is fast to compute and robust against small perspective changes, it fails to describe any spatial distribution of color within the image. To solve this, Huang et al. [29, 30] proposed the color correlogram, a color-based feature that integrates localized spatial information.

Let  $I$  be an image with the colors quantized into  $m$  colors  $c_1, \dots, c_m$ , and let  $I_c := \{p : I(p) = c\}$  be the set of all pixels in the image that have the color  $c$ . The color correlogram is then defined as:

$$\gamma_{c_i, c_j}^{(k)}(I) := \frac{P}{p_1 \in I_{c_i}, p_2 \in I} [p_2 \in I_{c_j} : ||p_1 - p_2|| = k]$$

It describes the probability that the color  $c_j$  can be found in a distance of  $k$  from a pixel of color  $c_i$ . One can restrict the correlogram to identical colors ( $c = c_i = c_j$ ) to obtain the autocorrelogram :

$$\begin{aligned} \alpha_c^{(k)}(I) &:= \gamma_{c,c}^{(k)}(I) \\ &= \frac{P}{p_1 \in I_c, p_2 \in I} [p_2 \in I_c : ||p_1 - p_2|| = k] \end{aligned}$$



In this thesis, a modified version proposed by Hofmann and Ali [36] is used, which computes the probability that the observed pixel  $p_1$  has surrounding pixels with the same color:

$$\alpha'_c(I) := \frac{P}{p_1 \in I_c, p_2 \in I} [p_2 \in I_c : p_2 \in \mathcal{N}(p_1)]$$

where the surrounding pixels  $\mathcal{N}(p_1)$  are defined by a mask with  $p_1$  as center.

Huang et al. [30] showed that the color correlogram itself is able to outperform histogram-based color features. To further improve its performance, the autocorrelogram is concatenated with a 300-dimensional color histogram in IHLS color space, resulting in a 600-dimensional feature [36]. The color correlogram features are computed using an in-house implementation.

### 3.2.2 Bag of Visual Words

Bag-of-words representations have their origin in text retrieval, where a document can easily be described by counting the occurrences of certain words within the document. An image can be described by a set of local patches extracted from various regions or points in the image, similar to the words within a text document. In contrast to words, however, they lack any meaningful ordering, and the number of different patches is potentially unlimited. So to be able to describe an image in a similar way to a text document, these patches have to be quantized into a limited *vocabulary* of visual words first.

To obtain a visual vocabulary, from which the visual words are then derived, a large set of local image patches is sampled from the dataset. In concept detection, these are often based on the local SIFT features introduced by Lowe [39]. These are then quantized using a k-means clustering algorithm (see Section 4.3), with the resulting cluster centers forming the visual vocabulary.

To now obtain a vector representation of an image, local patches are extracted from the image and matched against the vocabulary. Each patch in the image is thereby assigned to the closest cluster center, resulting in a description of the image by a set of visual words. By counting their number of occurrences in a histogram, the so-called bag of visual words representation is obtained [52]. The whole process can be seen in Figure 3.2.

While this feature regularly outperforms other features (for example, see [42]), it is computationally expensive to obtain. In addition to extracting thousands of patches, one also has to perform clustering on them and match all patches against the resulting cluster centers.

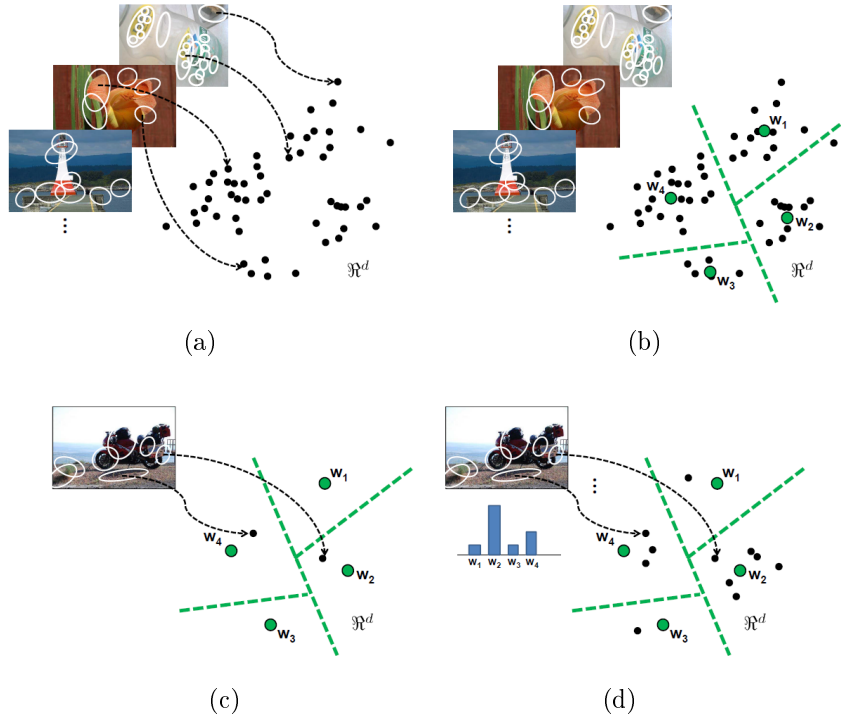


Figure 3.2: Obtaining bag of visual words features: (a) a large set of patches from various images is extracted and (b) clustered into a codebook using the k-means algorithm. (c) The bag of visual words representation for a new image is obtained by extracting the local patches and (d) matching them against the previously created codebook (Images taken from [22]).

In addition to using the bag-of-visual word features as explained above, probabilistic Latent Semantic Analysis (PLSA) [25] is performed to reduce the number of dimensions, as proposed by [52]. By doing this, both the time to train a classifier and the storage space needed can be reduced. This allows for an evaluation of the proposed system with regard to scalability concerns.

### 3.3 Classifiers

#### 3.3.1 Support Vector Machines

A Support Vector Machine (SVM) is a supervised machine learning method, often used for data mining tasks such as data classification. First introduced in their current (soft-margin) form by Cortes and Vapnik [13], they rapidly gained popularity due to a promising performance and a wide range of potential ap-

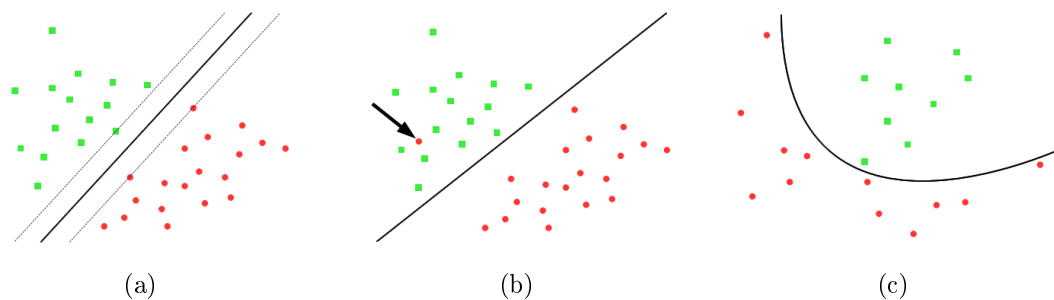


Figure 3.3: (a) maximum-margin hyperplane (b) soft-margin hyperplane, and (c) 2-dimensional data that is not linearly separable, but becomes separable in higher dimensions.

plications. Support Vector Machines have been applied successfully to a large number of problems, such as face recognition [34], spam recognition [17], but also in image and video retrieval [57]. This section introduces the basic principles of a Support Vector Machines, for a more detailed treatment see, for example, the book of Cristianini and Taylor [15].

Support Vector Machines are, in essence, binary linear classifiers. Given an unseen sample, the Support Vector Machine decides whether it belongs to one class or the other. To achieve this, an initial training step is required, where the SVM learns a statistical model based on a set of labeled training samples. Support Vector Machines can be characterised by three properties: a maximum-margin hyperplane, the soft-margin and the kernel functions.

### Maximum-margin Hyperplane

As a linear classifier, the model is described by the notion of a hyperplane, namely one that is able to separate the two classes. However, the number of candidate hyperplanes is infinite, which leads to the question what hyperplane makes for the best separation. Intuitively, one wants to separate both classes "as much as possible", thus by choosing a hyperplane that has the maximal distance to samples on both sides. This is expressed by the *maximum-margin* principle.

Given a set of labeled training samples  $x_i \in \mathbb{R}^d, i = 1, \dots, n$ , the hyperplane can be described by:

$$w \cdot x + b = 0$$

So a hyperplane that separates the given training samples with labels  $y_i \in \{1; -1\}$  (1 denotes a positive sample,  $-1$  a negative one) has to satisfy the following condition:

$$y_i(w^T x_i - b) \geq 1$$

The margin (see Figure 3.3 a) is defined as the distance from the hyperplane to the closest samples on both sides, and it can be shown that this distance equals  $\frac{1}{\|w\|}$  (see [23]). To now find the optimal hyperplane, this margin has to be maximized, resulting in the following optimization problem:

$$\begin{aligned} \underset{w,b}{max} \quad & \frac{1}{\|w\|} \\ \text{subject to} \quad & y_i(w^T x_i - b) \geq 1 \\ & i = 1, \dots, n \end{aligned}$$

As this is difficult to solve, it is usually rewritten to

$$\begin{aligned} \underset{w,b}{min} \quad & \frac{1}{2} \|w\|^2 \\ \text{subject to} \quad & y_i(w^T x_i - b) \geq 1 \\ & i = 1, \dots, n \end{aligned}$$

Maximizing  $\frac{1}{\|w\|}$  is equivalent to minimizing  $\frac{1}{2}\|w\|^2$ , so both expressions have the same solution. The second one, however, can be solved efficiently by Quadratic Programming [12].

### Soft-margin

It is often not possible to find a hyperplane, such that all samples from one class lie on the same side (see Figure 3.3 b). This can happen because the data is noisy or not fully linearly separable. A SVM should still be able to find a good separation in that situation, which is solved by allowing a certain number of errors. Formally, a soft-margin parameter  $\varepsilon \geq 0$  is introduced, which describes the degree of misclassification for a sample.

The problem now changes to:

$$\begin{aligned} \min_{w, \varepsilon, b} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \varepsilon_i \\ \text{subject to} \quad & y_i(w^T x_i - b) \geq 1 - \varepsilon_i \\ & i = 1, \dots, n \end{aligned}$$

The parameter  $C > 0$  regularizes this effect by penalizing misclassifications (non-negative  $\varepsilon$ ). The smaller  $C$ , the more misclassified samples are allowed and thus the margin grows.

### Kernel-trick

While the soft-margin allows some samples to be misclassified, the data might still not be linearly separable in the input space. Separability, however, can usually be achieved by transforming the data to a high dimensional feature space, where it becomes linearly separable again. To construct such a hyperplane in feature space, the samples have to be mapped using a transformation function  $x \rightarrow \Phi(x)$ .

The algorithmic solution of the SVM optimization problems requires the computation of inner products of the form  $\langle x_i, x_j \rangle$  [23]. So using a transformation, one would have to compute the inner product  $\langle \Phi(x_i), \Phi(x_j) \rangle$  in the high dimensional projected space, which is computationally expensive. The solution is the so called kernel-trick: instead of mapping the data using  $\Phi$  and calculating the inner products in the high-dimensional feature space, it is done in one operation using a kernel function:

$$K(x_i, x_j) \equiv \langle \Phi(x_i), \Phi(x_j) \rangle$$

Depending on this kernel function, the samples are mapped into different higher-dimensional spaces. Usually non-linear kernel functions are used, as most data is not linearly separable. In this thesis, two frequently applied kernel functions are used, the *RBF*-kernel (or Gaussian kernel)

$$K_{RBF}(x, y) := \exp[-\gamma \cdot \|x - y\|^2]$$

and the  $\chi^2$ -kernel

$$K_{\chi^2}(x, y) := \exp \left[ -\frac{1}{\gamma} \sum_i \frac{(x_i - y_i)^2}{(x_i + y_i)} \right]$$

where  $\gamma$  is called kernel parameter. Both kernels have been used successfully in concept detection before [35, 67].

Each SVM is described by two parameters, the cost parameter  $C$  and the kernel parameter  $\gamma$ , and their choice has a strong impact on the performance of the Support Vector Machine. Depending on the given problem, different values are necessary to obtain good results. Consequently, one has to identify good values for the two parameters before training a Support Vector Machine on a given problem. For this task, a grid search cross-validation is suggested [27]: different value-pairs are tried and the one with the best cross-validation result is picked. A second round is conducted by doing a finer grid-search around these two values, improving the previously selected parameters. However, it is suggested to do the parameter search only on a subset of the dataset, as it is a very time-consuming process.

In this thesis, the LIBSVM<sup>1</sup> implementation was used in the experiments.

### 3.3.2 PAMIR

Although Support Vector Machines often show an outstanding accuracy in concept detection systems, classification and learning becomes very slow with a growing number of samples. This is especially important when dealing with large-scale video databases (like YouTube and other web video portals) and a large semantic concept vocabulary. For this reason, a model with very fast learning capabilities should also be considered in the context of this thesis. Therefore, in addition to Support Vector Machines, the proposed system is also evaluated the fast, linear PAMIR-classifier.

The Passive-Aggressive Model for Image Retrieval (PAMIR) is a linear discriminative model originally proposed by Grangier et al. [21]. Initially designed for the retrieval of images from text queries, an alternative version was also applied to concept detection by Paredes et al. [44]. In this context, the algorithm tries to find a projection from the feature representations to the (one-dimensional) concept space, described by a weight vector  $w_c$ . Given a sample  $x_i$ , a score for the concept  $c$  then is obtained by calculating  $s(c, x_i) = w_c x_i$ .

---

<sup>1</sup><http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Mathematically, the task of finding such a weight vector can be described as maximizing

$$J(w_c) = \sum_{\forall x_p \in X_p} \sum_{\forall x_n \in X_n} (w_c x_p - w_c x_n)$$

where  $x_n \in X_n$  are negative samples, meaning the concept does not appear in it, while  $x_p \in X_p$  denotes a positive sample. However, for optimizing this term one would have to consider all  $|X_n| \cdot |X_p|$  possible combinations of  $x_p$  and  $x_n$ , resulting in a very expensive optimization task. To avoid this, an iterative solution based on the Passive-Aggressive algorithm [14] was proposed. It constructs a series of weight vectors  $w_0, \dots, w_n$  according to the following procedure (see [44]):

$$w_c^i = \underset{w_c}{\operatorname{argmin}} \frac{1}{2} \|w_c - w_c^{i-1}\|^2 + Cl(w_c; x_p, x_n)$$

where  $l$  is the hinge loss function, defined by

$$l(w_c; x_p, x_n) = \max(0, 1 - w_c(x_p - x_n))$$

In summary, this algorithms tries to minimize the total loss over a series of iterative updates of the weight vectors. The process is stopped after a predefined number of iterations, in general much lower than the number of possible sample pairs. Compared to the SVM, up to 10.000 times faster training times can be achieved [44]. An in-house Python-implementation was used for the experiments.





# Chapter 4

## Inferring Demographic Profiles by Clustering

### 4.1 Idea

Demographic attributes have been used as additional parameters in behavioural-based advertising systems [9], but it is also possible to target certain demographic groups directly. In fact, many advertisers define the target groups for their products and services in terms of demographic attributes like age, gender, education and income [26]. This approach offers an alternative to tracking a user's browsing and searching behaviour, which often raises security and privacy concerns.

However, the success of demographic targeting heavily depends on the definition of the demographic groups, as the perception of advertisements changes significantly between different ages and genders. It has been shown, for example, that the older the customers, the more they prefer images where the face of the subject is not cropped or hidden [48]. Consequently, targeting a broad demographics might lead to negative results, for example, when using the same advertising campaign for both children and adults. As a result, companies usually use more narrow definitions and tailor their campaigns accordingly (see Figure 4.1). In addition, many products are specifically targeted towards a narrow demographics (for example, acne creme for teenagers).

This leads to the question of how suitable demographic groups can be defined, for this thesis in the context a web video portal. One possibility is to define them manually, for example, by setting a specific age range for each of the groups ('male, 20 to 25 years old'). This, however, ignores the fact that many advertised products or services have a cross-over appeal. For instance, imagine a kind of sweets that is targeted towards teenagers, but also frequently bought by young

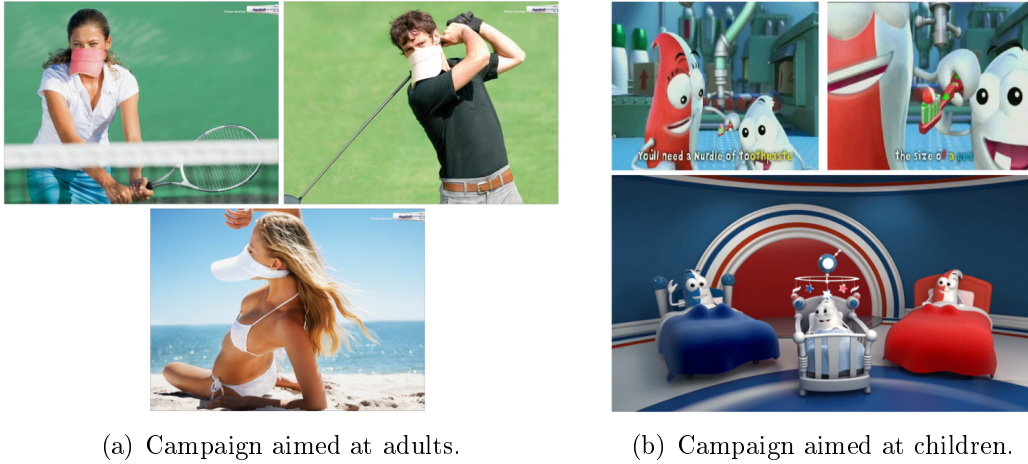


Figure 4.1: Example of how advertising strategies can differ depending on the target demographics. Both campaigns advertise tooth-paste from the same company, but one is targeted towards children and the other towards adults.

adults. In addition, we want to identify the groups that are characteristic to the observed environment, such that a further customization of the advertising process is possible. In the context of a web video portal, it is especially important to do this on basis of the videos the users watch.

To accomplish this, the idea of user clustering is adapted. It is a commonly used approach to detect groups of similar attributes and shared interests, and has been used successfully in the past, for example, in customer mining and behavioural targeting [60, 62]. However, instead of clustering the individual users, it is applied to the demographic profiles that describe the videos' audiences. How they are obtained will be explained in the next section.

An additional advantage of this approach is that clustering is an unsupervised process. This means that it can be applied even when no information about the user base is known beforehand, and it easily adapts to a different domain or platform (for example, to a large image database like Flickr<sup>1</sup>). In this case, the clustering can just be re-applied and thus adapted to communities with different age structures.

The result is a set of demographic clusters, each describing a group of videos sharing an age and gender distribution. Each of these clusters can then be associated with a set of suitable advertising campaigns, or the cluster membership of a user can be introduced as an additional, demographic parameter in an advertising system.

---

<sup>1</sup><http://www.flickr.com>



Figure 4.2: Common situation on a web video portal: the video has a lot more views than comments.

## 4.2 Demographic Profiles

Since the goal is to identify demographic clusters within the dataset, it is essential to decide on a way to describe the demographics of a videos' audience. One possibility is to aggregate the demographics of all users who watched the video, but detailed viewing statistics are usually not publicly available.

Instead, one can refer to information available through user comments on YouTube. It is reasonable to assume that the action of posting a comment on a video is a strong indicator that the author has watched it, and thus it can also be taken as a sign of interest in the video's topic. In the case of YouTube, these comments can be accessed via the YouTube Data API<sup>2</sup> and they always include the unique user name of the author. Thus it is possible to access the demographics of at least part of the audience through the public user profiles of the commenting users. However, it should be noted at this point that the number of comments is always substantially lower than the number of views (see Figure 4.2).

To describe now the demographic profile for a video, the age and gender of all users who posted a comment on this video is accumulated. This information is available for a large percentage (more than 80 %) of users, although some of them give age values that can be considered spurious. Most likely the main reason for this is the flagging of mature content on web video portals, which requires the user to be of full age to watch many videos. It is expected, though, that this has negligible impact on the performance of the clustering or the proposed demographics estimation.

<sup>2</sup><http://code.google.com/intl/de-DE/apis/youtube/overview.html>

After acquiring both age and gender information for a sufficient number of users, two age histograms are constructed: One for female users, and one for male users. As histogram bins the following age ranges are used, which are also applied in YouTube video statistics:

group	age range
teens	13-17
young adulthood	18-24
	25-34
middle adulthood	35-44
	45-54
late adulthood	55-64
	65-74
	75+

As you can see, the histogram bins are smaller for all users younger than 25 years. This is useful because more than half of the YouTube users falls into that range (see Table 6.1), thus the slightly higher resolution allows to capture more subtle differences for this group. In contrast, all users that are 75 years or older are counted in the same bin, as the number of users drops with increasing age. Children under 13 years are not taken into account.

For each video  $v$ , the two resulting 8-dimensional histograms  $h_{age}^{male}(v)$  and  $h_{age}^{female}(v)$  are first normalized and then concatenated to form the *demographic profile* for this video:

$$h_{age}(v) = h_{age}^{male}(v) \parallel h_{age}^{female}(v)$$

The whole process is illustrated in Figure 4.3.

However, to model the audiences' demographics accurately, a certain number of comments is required. Otherwise, the process would result in a very sparse histogram, which cannot be assigned reliably to a cluster. Empirically, the limit was set to 20 comments by different users with available age and gender information. This was found to be a good compromise between the quality of the resulting histograms and the amount of videos that had to be dropped (about 60 %).

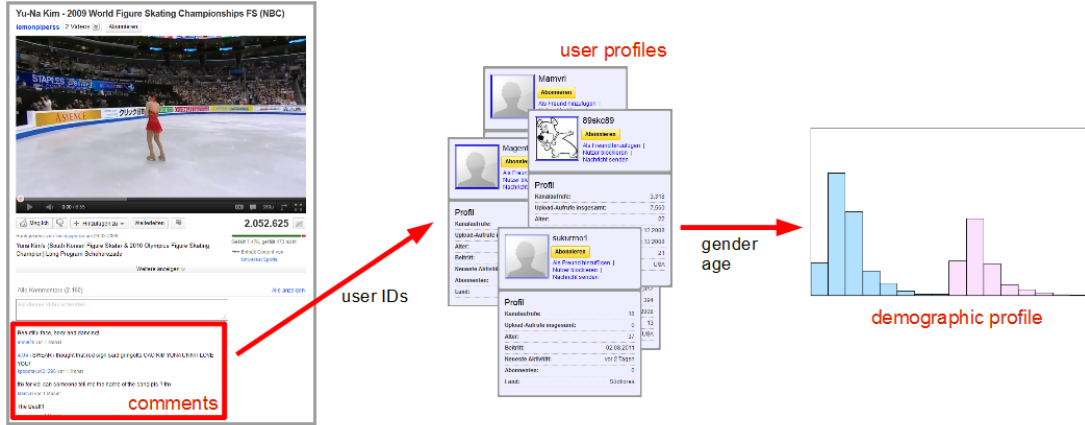


Figure 4.3: To create a demographic profile for a video, the unique user names from all users who commented on the video are collected. These are then used to access their public profiles, and the available age and gender information is aggregated to a histogram.

### 4.3 K-means

While originally proposed more than 50 years ago [56], k-means is still one of the most widely used clustering algorithms. The main reason for this popularity lies in its simplistic nature. It is easy to implement and can be applied effectively, even to large datasets. In this thesis, it is applied to create both the codebook for the bag-of-visual words representation and to model the demographics of the videos' audiences.

Formally, k-means tries to find a partition of all given samples into  $k$  clusters, such that the squared error over all clusters is minimized. With  $\mu_k$  being the mean of cluster  $c_k$  and  $X = \{x_i : i = 1, \dots, n\}$  the samples to be clustered, the squared error for cluster  $c_k$  is defined as

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

with  $x_i \in c_k$  being all samples assigned to cluster  $c_k$ , meaning that  $c_k$  has the closest mean to  $x_i$ . This leads to minimizing the following term:

$$J(C) = \min \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

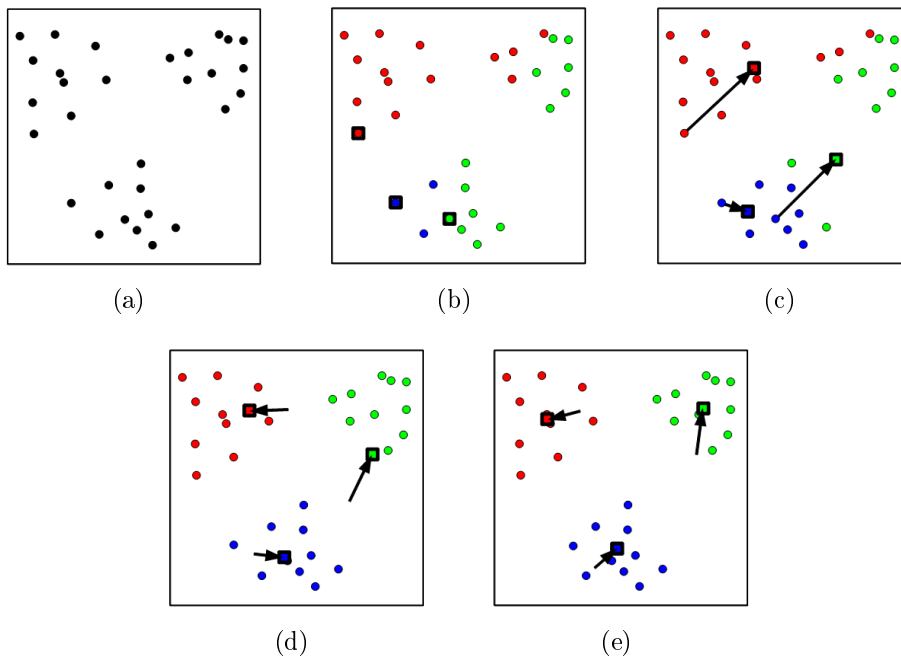


Figure 4.4: An example k-means run, the squares mark the cluster centers: (a) data to be clustered, (b) initial clusters, (c)(d) iterations, the cluster memberships change and cluster centers shift, (e) final iteration, the cluster memberships do not change.

The standard approach to solve this is as follows [31, 32]:

1. Select an initial set of  $k$  cluster centers  $\mu_1, \dots, \mu_k$
2. Assign each sample to the closest center
3. Recompute all cluster centers by shifting them according to the current cluster memberships
4. Repeat steps 2 and 3 until a convergence criteria or given number of iterations is reached

An example run can be seen in Figure 4.4.

The quality of the resulting clusters depends strongly on the initial selection of cluster centers, as k-means only converges to local minima. While multiple approaches were suggested so far, a random initialization has proven to be a good and effective choice [45]. One can also reduce this effect can by running the k-means algorithm several times and selecting the result with the lowest squared error.

## 4.4 Clustering Results

So far, this chapter has discussed how suitable demographic groups can be obtained as a basis for the demographics estimation on a web video portal. In this section, the results of the clustering on the dataset used in this thesis (see Chapter 6) are presented. The clustering was performed for different values of  $k \in \{5, 7, 10\}$ , and after a visual inspection of the resulting clusters,  $k$  was fixed to 7.

The results can be seen in Table 4.1, a visual impression in Figure 4.6. For each cluster, the center point (as a demographic histogram) and the most prominent semantic concepts (meaning the concepts with the largest number of videos in this cluster) are listed. The semantic concepts are based on the vocabulary that was used to download the videos from YouTube, and will also serve as a semantic vocabulary for the concept detection systems in this thesis. A complete listing of the vocabulary can be found in Table A.1.

Clusters 1 to 3 capture an almost exclusively male audience. The topics vary depending on the age, but there is also some cross-over appeal to be noticed. Clusters 2 and 3 are focused on ages 18 to 34, and in both concepts like 'boxing' or 'shooting' appear in top position.

Many of the female users seem to be interested in semantic topics like 'baby', 'dancing' or 'cake', which are mostly split between the concepts 4 and 5. However, cluster 5 describes a predominantly young female audience, with a strong focus on ages 24 or lower. In this age group, there seems to be a strong affection towards concepts like 'horse', 'anime' and 'cheerleading'. Cluster 4 on the other hand also features female users with an age of 25 or higher, and in addition a certain percentage of male users. Concepts like 'singing' or 'cooking' seem to attract both a male and female audience across different ages.

Cluster 6 seems to cover topics that are of interest to the average YouTube user, independent of their age. It covers widely popular concepts like the TV shows 'americas-got-talent' and 'muppets', but also 'commercial' and 'music-video', which both appeal to a broad demographics.

Finally, cluster 7 describes a mostly older and male audience, especially interested in topics that can be considered political. It features concepts for several politicians like 'obama' and 'mccain', but also the closely related concepts like 'press-conference' and 'interview'.

Overall, there is a strong connection between some concepts and the resulting demographic clusters. The concept detectors for these concepts are most likely strong indicators for the corresponding demographic cluster.

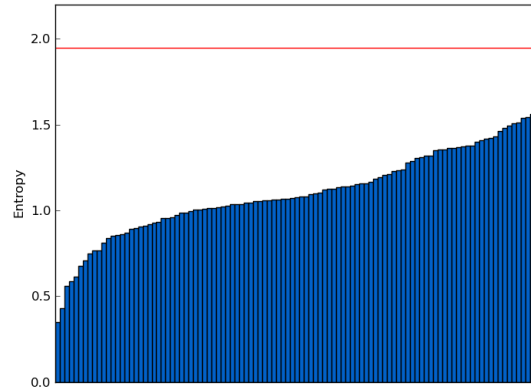


Figure 4.5: Entropy plotted for 105 semantic concepts. It measures the 'peakedness' of a distribution, with a low value indicating a strongly peaked distribution. The red line depicts a uniform distribution.

To get an estimate on how widely the concepts are distributed over the clusters, the 'peakedness' is measured for each concept's distribution, using the entropy as a measure.

The result can be seen in Figure 4.5, the entropy of a uniform distribution is included for comparison. As it turns out, the peakedness varies notably between the different concepts. No concept seems to be uniformly distributed, although some of them come close and thus these concepts make weaker indicators of demographic interest. On the other hand, some concepts seem to be strongly peaked, and thus are likely to contribute positively to the demographics estimation. To get a better impression of the impact on the demographics estimation, this aspect will be evaluated in a later experiment.



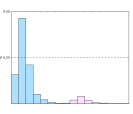
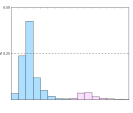
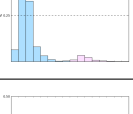

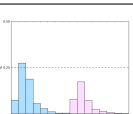
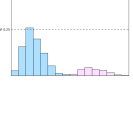

cluster nr.	histogram	top concepts
1		counterstrike-game, skateboarding, worldofwarcraft, darth-vader, simpsons, soccer, dark-skinned-people, breakdancing, windows-desktop, graffiti
2		mccain, boxing, interview, talkshow, riot, georgew-bush, poker, shooting, obama, soldiers
3		crash, boxing, race, car, worldofwarcraft, phone, shooting, wheel, basketball, soccer
4		singing, cake, cooking, choir, food, baby, kitchen, cats, dancing, dogs
5		horse, anime, cheerleading, kiss, gymnastics, cake, riding, dancing, videoblog, baby
6		americas-got-talent, cats, cartoon, origami, piano, muppets, commercial, tornado, choir, music-video
7		obama, mccain, georgewbush, court, interview, press-conference, airplane-flying, riot, demonstration, concert

Table 4.1: Resulting clusters for  $k=7$  with the most prominent semantic concepts in each cluster.



(a) **Cluster 1** has a strong focus on male teens, covering youth activities like skateboarding and breakdancing, but also video games.



(b) **Cluster 2** is oriented towards middle-aged adults. Many political concepts appear, the activities are more 'mainstream'.



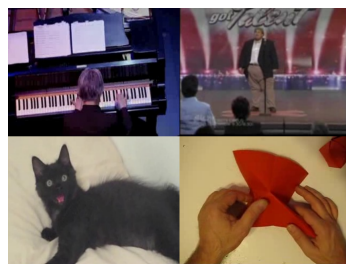
(c) **Cluster 3** covers young adults. Technical gimmicks like smartphones and cars play an important role, more sports activities appear.



(d) **Cluster 4** is the first cluster with a high rate of female users. Many food-related concepts can be found, but also social activities like 'singing' and 'dancing'.



(e) **Cluster 5** focuses on young females, which show a strong interest in activities like 'cheerleading' and 'horse riding'. Many concepts from cluster 4 also appear here, although less prominent.



(f) **Cluster 6** is the most universal cluster, covering a broad audience and popular topics like casting shows and animal videos.



(g) **Cluster 7** is the 'oldest' cluster, political concepts like 'obama' as well as news-related concepts like 'interview' or 'press-conference' clearly dominate.

Figure 4.6: Visual impressions for seven demographic clusters.

# Chapter 5

## Linking Video Content with Demographics

### 5.1 Overview

Chapter 4 discussed an approach to identify demographic groups within a large set of videos from a web video portal. The results have shown that there is a strong connection between the identified demographic clusters and the semantic concepts the videos focus on. Obviously, this relationship could form the basis for a system to estimate the demographics of a videos' audience. On the other hand, available demographic information could also be used to improve concept detection performance.

In this chapter, several systems will be introduced that implement these ideas. Section 5.2 describes three approaches to estimate what demographic cluster a videos' audience belongs to, while Section 5.3 details how the demographic profiles can be used as a new modality in a state-of-the-art visual concept detection system.

### 5.2 Demographics Estimation by Concept Detection

In the following, several methods for predicting the demographics of a video's audience are proposed, which only rely on low-level visual features extracted from the video stream. This makes it possible to estimate demographic parameters even when no textual information is available or when the identity of the user is unknown. For example, these methods can be applied to recently uploaded

videos, which often lack a meaningful description or comments. In addition, it does not require tracking the users' searching and browsing behaviour, avoiding any adherent privacy concerns. Another imaginable application is the search for specific scenes within a video that are of interest to a given demographics. For example, think of a news program that consists of many different reports.

In the following, it is assumed that a set of demographic cluster  $D$  has already be identified by the initial training step. The next step requires the extraction of low-level visual features (see Section 3.2), which form the basis for the following approaches:

- A **marginalization-based** approach, where first a set of semantic concept detectors is trained. The demographics estimate is then obtained from their results and the demographic cluster model by marginalization over the concepts.
- A **two-step** approach, where a set of semantic concept detectors is used to create a new intermediate feature representation. This feature is then used to train a second set of classifiers, now for estimating the demographics.
- A **baseline** approach, which trains a set of supervised classifiers to estimate the demographics directly on the visual features.

The outputs for all three approaches are posterior probabilities  $P(d|x)$ , denoting the probability of a video - described by the feature representation  $x$  - belonging to the demographic groups  $d \in D$ .

### 5.2.1 Baseline System

As seen in Figure 4.6, there is a high diversity in visual appearance between the different demographic clusters. However, it is still possible that within a cluster the videos share similar visual characteristics. For example, as mentioned before, older people prefer seeing full faces in contrast to cropped faces, and also the perception of color is influenced by the age [48].

This proposed baseline system tries to make use of these potential differences by directly learning a classifier for each of the identified demographic clusters. This shows some similarities to concept detection, if the demographical clusters are interpreted as concepts ('cluster 1', 'cluster 2', and so on). For each of the demographic clusters, a visual classifier is trained as described for a concept detection system (see Figure 3.1). An overview can be seen in Figure 5.1.

While the demographic clusters are visually more complex than semantic concepts like 'soccer' or 'dogs', the computational effort is considerably lower. The number

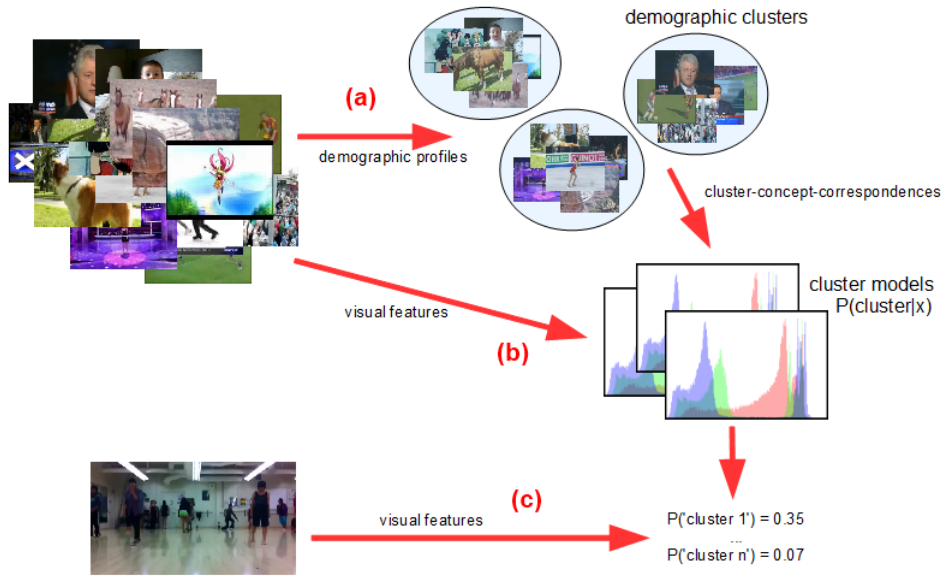


Figure 5.1: Overview of the baseline system. (a) A set of training videos is first clustered according to their demographic profiles. (b) Then for each of the resulting  $k$  clusters a concept detector is trained over visual features. (c) The demographics of a new video is estimated by extracting the visual features from a set of representative keyframes, and applying the  $k$  concept detectors.

of clusters used is quite small (here  $k = 7$ ), compared to the large number of concepts in a state-of-the art concept detection system.

The concept detectors are trained over different visual features, namely color correlograms, visual words and PLSA-reduced visual words. Two different machine learning algorithms are applied, a fast linear PAMIR classifier and Support Vector Machines. The output is a posterior  $P(d|x)$ , which describes the probability that a video with feature representation  $x$  has an audience described by the demographic cluster  $d$ .

## 5.2.2 Marginalization-based Approach

While the baseline approach is valid in principal, the complex visual nature of the demographic clusters created might not be captured sufficiently by the visual features representations. So the idea behind the marginalization-based approach is to divide this task into two parts:

1. A visual classification based on a large semantic vocabulary  $C$ .
2. Modelling the demographics distribution for these semantic concepts.

As before, the demographic profiles will be clustered to obtain a simple demographic model. However, instead of using the cluster membership to derive complex visual-demographic classes, this information is used to model how the initially defined semantic concepts are distributed amongst the demographic clusters. More specifically, for each concept  $c \in C$

$$P(d|c) = \frac{|\{v : v \in V_c \wedge f(v) = d\}|}{|V_c|}$$

is obtained, where  $d \in D$  is a demographic cluster, and  $V_c$  is the set of all videos that show the concept  $c$ . The function  $f : V \rightarrow D$  maps each video to a demographic cluster in  $D$ .

Afterwards, for each of the concepts in  $C$  a concept detector is learned, using the same visual features and classifiers as mentioned before. This yields a statistical model, which describes the posterior  $P(c|x)$ , the probability that a video with feature representation  $x$  shows a semantic concept  $c \in C$ . As before, the goal is to obtain an estimate for the membership of a video for each of the demographic groups  $d \in D$ , namely the posterior  $P(d|x)$ . This is obtained by marginalization:

$$\begin{aligned} P(d|x) &= \sum_{c \in C} P(d, c|x) \\ &= \sum_{c \in C} P(d|x, c) \cdot P(c|x) \\ &\approx \sum_{c \in C} P(d|c) \cdot P(c|x) \end{aligned}$$

An overview can be found in Figure 5.1. While this approach is more time-consuming, as the number of classifiers is a lot higher compared to the baseline system, the separation of the visual and demographic models is expected to outperform the baseline system. This is because the individual semantic concepts are visually a lot less complex than the mixture of semantic concepts found within one demographic cluster.

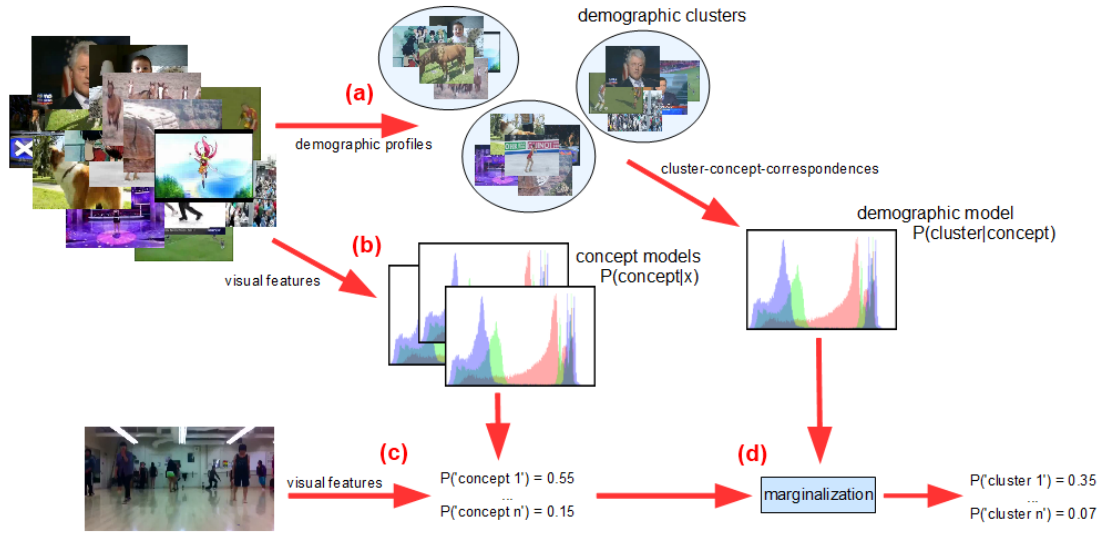


Figure 5.2: Overview of the marginalization-based system. (a) A set of training videos is first clustered according to their demographic profiles, which forms the basis of a simple demographics model. (b) Visual features extracted from the videos’ keyframes are used to train a set of concept detectors. (c) To map a new video to a certain demographics, visual features are extracted from a representative set of keyframes and matched against the concept models. (d) These scores are then used with the demographics model to form a prediction.

### 5.2.3 The Two-step Approach

The third approach suggested is similar to a technique called supervised classifier combination, where the outputs of several classifiers (for different features) are combined to train a new classifier. One advantage of this method that might also help linking visual features and demographics, is its ability to exploit non-linear relationships between the different modalities.

First, a set of  $n = |C|$  concept detectors is learned on a vocabulary of semantic concepts  $C$ , similar to the previous approach. However, their outputs  $P(c|x)$  are now combined to form a new intermediate feature representation:

$$\begin{aligned}
 feat_{2step}(x) &= P(c_1|x) \parallel P(c_2|x) \parallel \dots \parallel P(c_n|x) \\
 &= \prod_{i=1}^n P(c_i|x)
 \end{aligned}$$

These 'two-step' feature vectors are then used to train a second set of classifiers, one for each of the demographic clusters, similar to the baseline approach. The demographics of a new video is predicted by extracting the visual features from a set of representative keyframes, and matching them against the semantic concept models to create the two-step feature vectors. These are then matched against the second model to form the prediction. Figure 5.3 shows this process in detail.

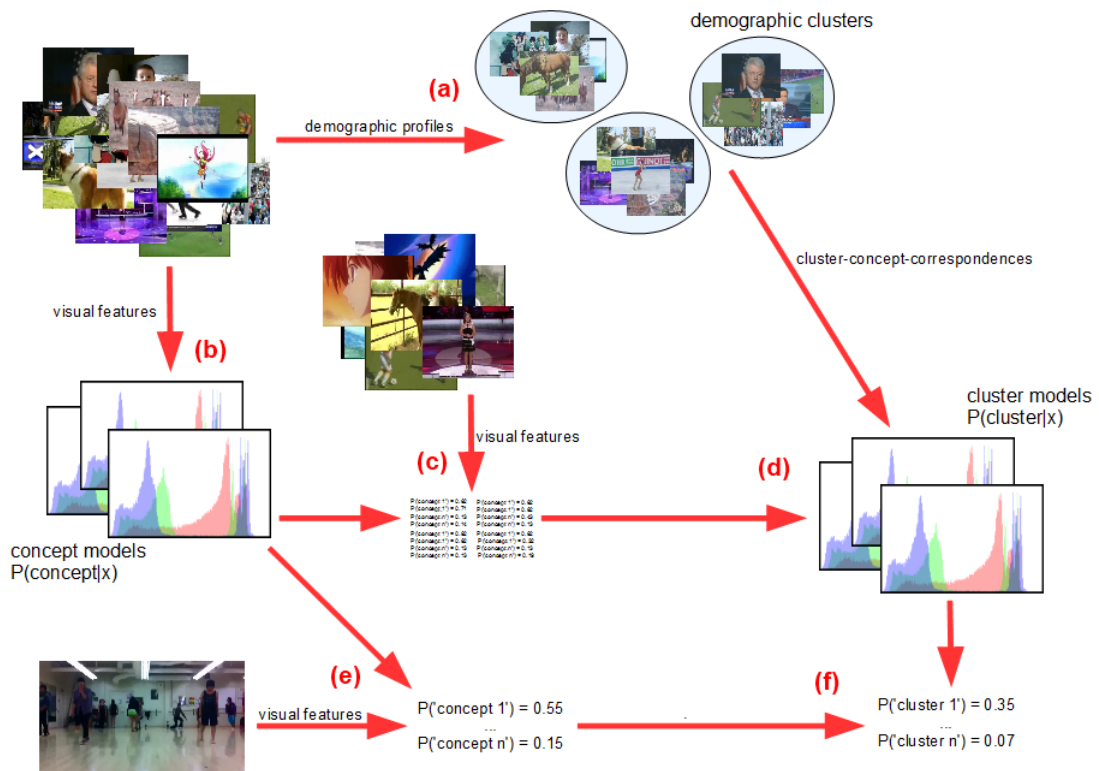


Figure 5.3: Overview of the two-step approach. (a) First, a set of training videos is clustered according to their demographic profiles. (b) Visual features are extracted from the same training set to learn a set of concept detectors. (c) Visual features from a second training set are matched against these concept models, and the results are used for creating the two-step feature representations. (d) These are then used to train a second set of visual classifiers, now for the demographic clusters. (e) For a new sample, the visual features are extracted and matched against the previously trained concept detectors, obtaining a two-step feature representation. (f) This is then used to obtain the demographics estimate by applying the cluster models.



## 5.3 Demographic Profiles for Improving Concept Detection

So far, three approaches were introduced that use the visual content of a video to estimate the demographics of its audience. Conversely, if the demographics for a video's audience are known (meaning a video is sufficiently commented), this information could be used for improving semantic concept detection. With better concept detection results, the retrieval quality of a web video search and the semantic indexing of unannotated videos can be improved considerably.

Therefore, this thesis proposes to use the demographic profiles - where available - as an additional feature for the concept detection system. While the combination of multiple features is an often applied principle in modern concept detection systems [54], the inclusion of viewer demographics as an additional, social signal can be considered a novelty.

The feature combination can be achieved by either a fusion of the features ('early fusion'), or a fusion of the classifiers ('late fusion'). Both approaches are applied in this thesis, and will be introduced in the following.

### 5.3.1 Early Fusion

By fusing different feature extraction results, one creates a new representation of the observed entity, in our case an image or a video. As every entity is now described by a single feature vector, regardless of the number of features combined, only one learning step is required. Here, two simple methods to combine the different features are proposed:

- **Vector concatenation**, where a new feature representation is obtained by concatenation.
- **Vector multiplication**, where a new feature representation is created by the outer product of the original feature vectors.

However, using these methods (especially the multiplication scheme) will increase the dimensionality of the feature used for classification, and thus invoking the so-called curse of dimensionality. It describes the exponential growth in computational complexity and a possible loss in generalization ability that comes with an increasing number of dimensions. To keep these effects to a minimum, the 2000-dimensional visual word features are reduced to 80 dimensions (see Section 3.2.2) before using them in an early feature fusion. For the same reason, color correlograms are only used in late fusion.

## Vector concatenation

Given a set of feature representations  $x_1(I), \dots, x_n(I)$  for a keyframe  $I$ , a new feature vector  $x_{conc}(I)$  is obtained by concatenation:

$$x_{conc}(I) = \left\| \begin{array}{c} x_1(I) \\ \vdots \\ x_n(I) \end{array} \right\|$$

Vector concatenation is a simple method for fusing different feature vectors, but it has shown good results in concept detection systems [53].

## Vector multiplication

Given two feature vectors  $x(I)$  and  $y(I)$  for a keyframe  $I$ , a new feature vector  $x_{\otimes}$  is obtained by calculating the outer product of both vectors:

$$x_{\otimes} = x \otimes y = \begin{bmatrix} x_1y_1 & x_1y_2 & \dots & x_1y_n \\ x_2y_1 & x_2y_2 & \dots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \dots & x_my_n \end{bmatrix}$$

with  $n$  and  $m$  being the number of dimensions of the two feature vectors to be fused. The resulting  $n \times m$  matrix is interpreted as a vector of dimension  $n \cdot m$ .

### 5.3.2 Late Fusion

Combining feature vectors is one of two possible ways to use different modalities in a concept detection system, the other way is to combine the classifier outputs itself. More specifically, a classifier is trained for each feature type and then the classifier scores are combined. As with feature fusion, there are different ways to achieve this:

- An **unsupervised combination**, which uses arithmetic operations like the average to combine the different results.
- A **supervised combination**, where a new classifier is trained over the results obtained by the single-feature classifiers.

In the unsupervised approach, the results of multiple classifiers (one per feature used) are combined using an arithmetic operation. For example, given the classifier results  $P_i(c|v)$  for a set of features  $\{f_i : i = 1, \dots, m\}$ , one can compute the average:

$$P_{avg}(c|v) = \frac{1}{m} \cdot \sum_{i=1}^m P_i(c|v)$$

Another often used method in this context is the maximum:

$$P_{max}(c|v) = \max_{i=1, \dots, m} P_i(c|v)$$

or the combination of both:

$$P_{avgmax}(c|v) = \frac{1}{m} \cdot \sum_{i=1}^m P_i(c|v) + \max_{i=1, \dots, m} P_i(c|v)$$

In this thesis, only an unsupervised combination is used. For more details on the supervised combination, see [54].



# Chapter 6

## Experiments

### 6.1 Dataset

The foundation for the following experiments is a set of approximately 30,000 videos with a total runtime of over 2,800 hours. The videos were automatically downloaded from the video portal YouTube, using the publicly available YouTube Data API<sup>1</sup>. They are spread over 230 semantic concepts, and for each concept an appropriate query was manually formulated (for example, 'skiing -water' for the concept 'skiing') and used to download approximately 150 videos downloaded per concept. A complete listing of the concepts and the queries that were used in the download process can be found in Table A.1.

From these videos, about 1,1 million keyframes were extracted using the adaptive clustering approach described in [8]. The resolution of the videos and keyframes is 320 x 240 pixels.

In addition, for each concept the comment feeds of 500 videos (including the downloaded videos) were retrieved using the same queries, together with the unique user names of the comment authors and the demographic parameters accessible through their user profiles.

This resulted in a total of 2,2 million unique users, of which more than 80 % gave their gender and an age between 13 and 95 years. A detailed breakdown of the users according to their age and gender is shown in Table 6.1. Notably, most of the users are male (75.44 %) and the majority is younger than 25 years old (51,19 %).

From this main dataset, several smaller sets were created for the different experiments as described in the following.

---

<sup>1</sup><http://code.google.com/intl/de-DE/apis/youtube/overview.html>

age range	% male	% female
13-17	6.95	4.65
18-24	29.43	10.16
25-34	24.90	5.49
35-44	8.06	2.13
45-54	3.90	1.31
55-64	1.49	0.60
65-74	0.40	0.13
75+	0.75	0.25
total	75.44	24.56

Table 6.1: Age and gender distribution in the YouTube dataset.

The first experiment evaluates if there is a change in concept detection accuracy, when splitting the data by the number of comments, instead of splitting it randomly. This is done on a subset of 10 randomly selected concepts, namely 'clouds', 'eiffeltower', 'hiking', 'orchestra', 'skiing', 'soccer', 'swimming', 'tony-blair', 'videoblog' and 'world-of-warcraft'. For each of these concepts, the following sets are created:

**VAL-RANDOM:** All available videos are randomly split into 5 complementary subsets, which are used in a 5-fold cross-validation. Keyframes sampled from one subset are used for testing, while keyframes from the other four subsets are used for training. This is repeated four times, such that each subset is used once for testing. Results are averaged over all five rounds.

**VAL-SPLIT:** For each concept, the available videos are ranked according to the number of unique users that commented on them. The training sets are created by sampling keyframes from the lowest ranked videos, until a given number of positive and negative samples is reached. Analogously, the test sets are created by sampling from the highest ranked videos.

The next sets are used in most other experiments, including the demographics estimation experiments. Therefore, it is necessary that all test videos can be reliably assigned to a demographic cluster, meaning that a minimum number of unique users commented on them. However, many of the concepts could not provide a sufficient amount of such videos, so the semantic vocabulary was reduced from 230 to 105 (see Table A.1).

**FINAL:** For each concept, the videos are ranked according to the number of unique users that commented on them, similar to VAL-SPLIT. The test set is created by randomly sampling 20 keyframes from each of the 50 highest ranked videos per concept. The training sets (one per semantic concept) are created by

sampling keyframes from all videos ranked lower, until a set number of positive (up to 5000) and negative samples (5 times the amount of positive samples) in each training set is reached.

To learn the baseline system, all videos used for training need a sufficient number of comments, so that they can be assigned to a demographic cluster. To achieve this, the previous set FINAL is modified:

**BASELINE:** For each semantic concept, the 10 most commented videos from the training sets in FINAL are selected. The BASELINE training sets (one for each demographic cluster) are created by first matching these videos to the demographic clusters, and then randomly sampling keyframes from them, according to their cluster membership. The number of positive keyframes in each training set is slightly increased (from 5000 to 6000), the negative keyframes are again sampled until a 1:5 ratio (positive to negative samples) is reached. The test set remains the same.

Obviously, the total number of videos used in the training sets of BASELINE is lower than in FINAL. However, as the number of training sets is significantly reduced, they still contain a sufficient number of samples, covering all semantic concepts.

## 6.2 Experiments

### 6.2.1 Concept detection

#### Experiment 1 - Validity of dataset split

To evaluate the proposed methods for demographics estimation reliably, a good ground truth on the test set is necessary. This, however, is only possible if all videos in the test set have a sufficient number of comments. If only a few users commented on a video, its demographic profile is sparse and thus the estimate of its true age structure is poor. Consequently, it has a negative impact on the evaluation. Therefore, the dataset will be split according to the number of comments for the later experiments: Videos with many comments are used for the evaluation, while videos with fewer comments are used in training.

However, it is possible that there are notable differences in the visual appearance between frequently commented and uncommented videos. For example, a professionally edited video often attracts more viewers than an amateur video. This

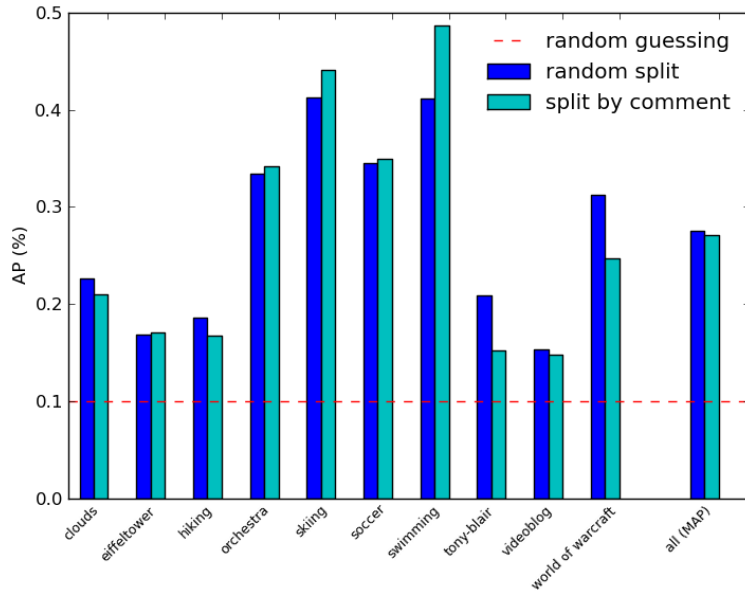


Figure 6.1: Concept detection results comparing a random split with a split by the number of comments. While for some of the concepts the average precision varies depending on the split strategy, the overall performance is comparable.

difference would also be captured in the extracted visual features and could introduce bias to the concept detection system, if splitting the data by the number of comments.

The goal of this experiment is now to validate that the chosen split strategy does not influence concept detection performance negatively. Two groups of concept detectors are trained for a randomly selected subset of 10 concepts, using the fast PAMIR classifier over 2000-dimensional visual words. The first group uses the randomly created set VAL-RANDOM in a 5-fold cross-validation. The second group uses the set VAL-SPLIT, for which the videos were separated into training and test set based on the number of users that commented on them.

The results can be seen in Figure 6.1. The first thing to be noticed is that some concepts show a bigger difference in average precision between the two split strategies than others. Concept detectors for the concepts 'world of warcraft' and 'tony blair' perform better when the split is done randomly, whereas for the concept 'swimming' it performs worse. This indicates that there is a link between the visual appearance and the comment frequency, although only for some concepts.

Overall, the performance of both systems is comparable, so the by-comment split seems to be a viable option for the following experiments. Its impact on the evaluation is expected to be negligible.



	SVM	PAMIR
correlograms	5.0	3.4
viswords(2000)	8.8	5.5
viswords(80)	6.3	3.9
random	0.9	

Table 6.2: Concept detection results (% mean average precision) for different combinations of classifier and feature. SVMs clearly outperform the simple PAMIR model, while the high-dimensional visual words prove to be the best feature.

## Experiment 2 - Concept detection

In this experiment, the concept detection accuracy for different combinations of classifiers and visual features is evaluated. The classifiers applied are the linear PAMIR model and Support Vector Machines with  $\chi^2$ -kernels. Color correlograms, 2000-dimensional visual words and 80-dimensional PLSA-reduced visual words are used as features. For training and testing the set FINAL is used, which is based on a by-comment split and 105 semantic concepts. The overall results can be found in Table 6.2, for detailed results on concept level see Table A.2.

The combination of Support Vector Machine and high-dimensional visual words achieves the best results with a mean average precision of 8.8 %, a 10-fold increase over random guessing (0.9 %).

The 2000-dimensional visual words outperform all other features, followed by the 80-dimensional visual words and the color correlograms. Although their mean average precision is more than 2 % lower, both features are still a suitable options in a large scale environment: The 80-dimensional PLSA features can be stored efficiently and allow for a faster classifier training and application, whereas color correlograms can be extracted in a short amount of time.

It can also be seen that the Support Vector Machine outperforms the simple linear model on every feature. However, the speedup that PAMIR achieves over the SVM is significant [44]. Again, this might be an important factor to consider for a web video portal, as the amount of videos that need to be processed is tremendous.

Overall, the results are within the set expectations. It is comparable to the improvements achieved in other benchmarks with a similar amount of concepts, like TRECVID [7]. Especially the combination of visual words and Support Vector Machine is a suitable candidate to use in the demographics estimations experiments.

	% <i>MAP</i>
demo-PAMIR	3.9
demo-SVM	6.5
viswords(2000)-SVM	8.8

Table 6.3: Concept detection results using demographic profiles as a feature and two different classifiers. The best visual words-based system serves as reference.

### Experiment 3 - Demographic context in concept detection

As already discussed before, there seems to be a strong connection between different demographic groups and the semantic concepts appearing in videos they watch. Intuitively, the demographic profiles obtained from videos (as explained in Chapter 4) could serve as a discriminative feature for concept detection.

Therefore, this experiment investigates the performance of a concept detection system that uses the 16-dimensional demographic profiles as a feature. PAMIR and Support Vector Machines are again used as classifiers and applied to the same training and test set as in Experiment 2 (FINAL).

The results can be seen in Table 6.3. The combination of Support Vector Machine and demographic profiles achieves a mean average precision of 6.5 %. This is not only a good improvement over random guessing (0.9 %), but also outperforms most of the other classifier-feature combinations from Experiment 2. Only the SVM trained over the high-dimensional visual words achieves a better result.

This is a surprisingly good result: By simply inspecting a video’s demographic profile, it is possible to achieve a concept detection accuracy comparable to content analysis. Even more so considering the low dimensionality of the demographic profiles and the negligible computational effort required to obtain them. However, there is a large difference in average precision between different semantic concepts (see Table 6.4). For example, the concept ‘cake’ achieves a very high average precision of 35.2 % with the demographic profiles, indicating that this concept is closely tied to a specific demographic group. On the other hand, some concepts (for example ‘ice-skating’) perform significantly better with visual words. This could indicate that these concepts attract a broader demographics, so that the demographic profile loses its discriminative power. Another possibility is that the number of comments available was too low to learn a good model.

Either way, the results confirm the observation that was already made in Chapter 4, namely that many semantic concepts have a specific audience in terms of demographic attributes. In addition, the fact that one feature is often clearly better than the other supports the idea to combine both for a more robust concept detection. This will be evaluated in the next experiment.

concept	viswords(2000)-SVM	demo-SVM	$\Delta$
cake	8.1	35.2	<b>27.1</b>
counterstrike-game	12.1	36.7	<b>24.5</b>
riding	9.0	30.6	<b>21.6</b>
horse	7.0	24.4	<b>17.3</b>
baby	2.5	18.5	<b>16.0</b>
...	...	...	...
boxing	32.1	12.2	-19.9
tennis	21.5	1.6	-19.9
rugby	21.8	1.7	-20.1
origami	32.0	6.4	-25.6
ice-skating	33.8	5.3	-28.5
all	8.8	6.5	-23.0

Table 6.4: List of the semantic concepts that show the highest and lowest difference in average precision, comparing visual words and demographic profiles on a SVM.

#### Experiment 4 - Combining visual features and demographic context:

So far it could be seen that a concept detection based on visual features works well for some concepts (like 'ice-skating' or 'origami'), while others (for example 'horse' or 'baby') show a far better performance using demographic context. This indicates that using both modalities together might lead to a significant improvement in concept detection accuracy.

Therefore, concept detection systems that use both visual features and the demographic context are evaluated in the following. To accomplish this, the early- and late-fusion strategies explained in Section 5.3 are applied. As before, FINAL serves as training and test set.

Because of the high number of possible combinations of features and classifiers, not all are evaluated, notably missing the early fusion of demographic profiles with color correlograms and high-dimensional visual words. Since they have shown a superior performance so far, Support Vector Machines are used as classifiers.

The results can be found in Table 6.5. Regardless of the fusion method and the visual feature, the additional use of demographic context leads to a strong improvement in concept detection performance of up to 4.1 %, which is a relative improvement of nearly 50 %. As before, the use of visual words still yields the highest average precision, but the difference to the dimensionality-reduced version shrinks from 2.5 % to 1.3 %.

The concatenation of the 80-dimensional visual words feature with the demo-

	early fusion		late fusion	
	concatenate	multiply	average	avgmax
correlograms	-	-	9.5	8.9
viswords(80)	11.6	10.3	10.8	10.6
viswords(2000)	-	-	12.9	12.5

Table 6.5: Concept detection results (% MAP) for the fusion of visual and demographic features.

graphic profiles achieves a mean average precision of 11.6 %, still outperforming the best single-feature concept detection system by 2,8 %. In addition, since this combined feature representation has only 96 dimensions (compared to 2000), the learning time and storage complexity are reduced considerably.

In Figure 6.2 one can see the top results for the concepts 'surfing' and 'poker', for which a particularly strong improvement was observed. The visual words alone performs reasonably well for both concepts, but tend to cause many false positives. For example, the concept detector for 'surfing' also reacts to videos showing generic beach scenes, and the detector for 'poker' gives high scores to all scenes that show close-ups of persons. This is improved by using the demographic context, removing many false positive results.

In summary, the use of demographic context in concept detection leads to a significant improvement in average precision. However, it should be noted that this approach can only be applied in a domain where demographic information is available. This automatically limits it to videos with a minimum number of views or, in our case, comments.

## 6.2.2 Demographics estimation

### Experiment 5 - Demographics estimation baseline

The results of Experiment 3 have shown that semantic concepts can be strong indicators of demographic interest. This connection was used in Experiment 4 to improve concept detection performance considerably, by fusing visual features with the demographic profiles introduced earlier. In the following, the other direction is considered, namely the estimation of demographics based on visual concept detection alone.

It is possible that the visual differences between the demographic groups are significant enough, so that it is not necessary to rely on specific concept detectors for the individual semantic concepts used so far. To investigate this further, as a

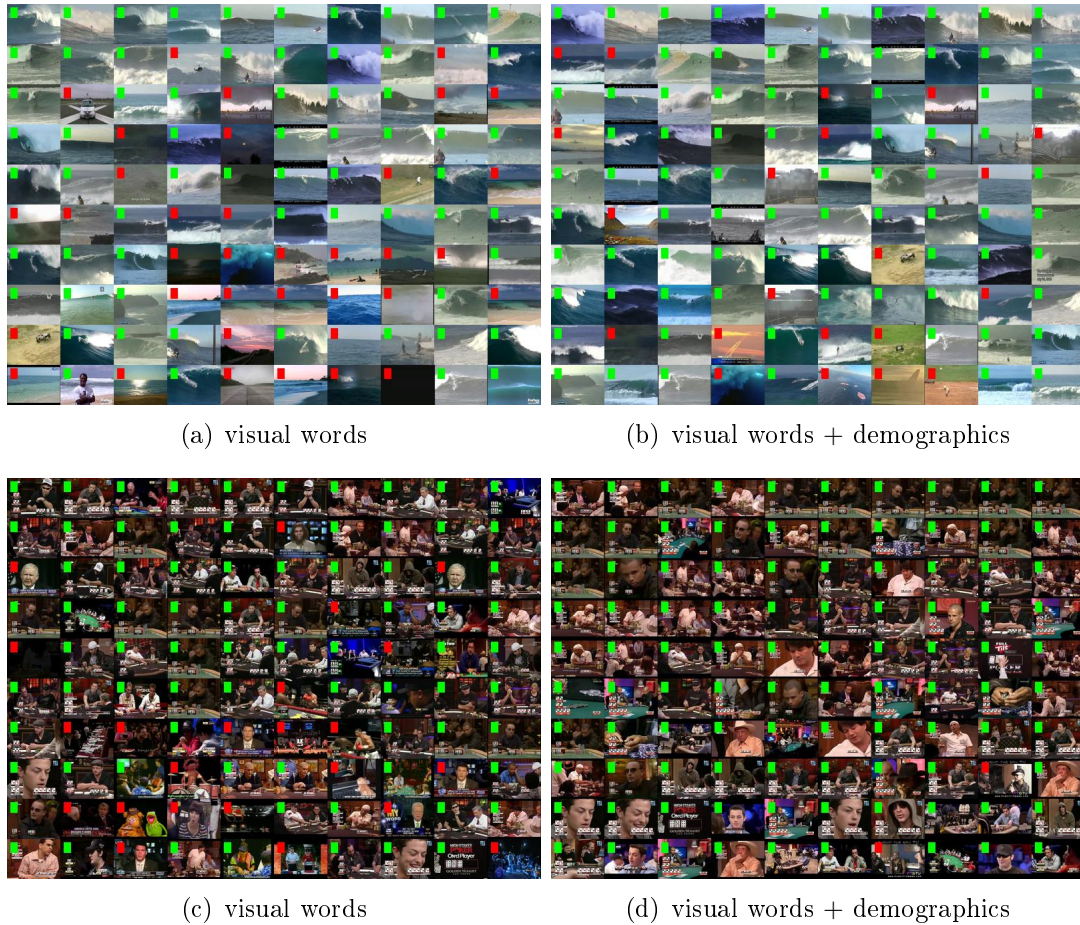
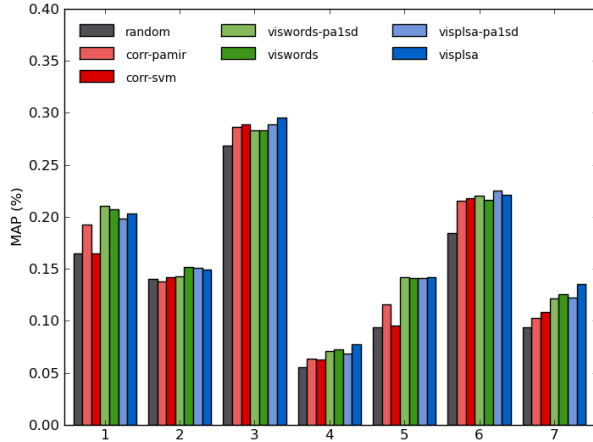


Figure 6.2: Top rated keyframes for two concepts ('surfing', 'poker') where the use of demographic context helps improving concept detection results. Red blocks indicate false positives, green blocks correctly identified keyframes. In both cases, using the demographic profiles removes many false positives that showed similar visual scenes.

simple baseline, the demographic groups are learned directly over visual features, without relying on the previously trained concept detection systems.

The resulting system shows some similarities to a concept detection system, if the demographic clusters are interpreted as semantics concepts ('demographic cluster 0', 'demographic cluster 1', and so on). For each of them a Support Vector Machine and a PAMIR model are learned, using BASELINE for training and testing. The results are displayed in Figure 6.3.

The combination of Support Vector Machines and 80-dimensional visual words could achieve the highest mean average precision of 17.5 %. However, given a mean average precision of 14.3 % for random guessing, this is a fairly low increase.



(a)

	SVM	PAMIR
correlograms	15.4	15.9
viswords(2000)	17.1	17.0
viswords(80)	17.5	17.1
random	14.3	

(b)

Figure 6.3: Demographics estimation results (% MAP) for the baseline system. (a) shows the detailed results on cluster level, (b) the overall results.

This outcome was expected to some degree, as the seven demographic clusters are visually much more complex than the 105 semantic concepts used so far. In some demographic clusters, several concepts appear that have a very different visual appearance, like 'horses' and 'cheerleading' (see Figure 6.4).

Also, many concepts share visual features, like the green in 'soccer' and 'golf', but do not attract the same audience. 'Golf' is stronger oriented towards middle-aged to late-aged adults (demographic clusters 2 and 7), while 'soccer' is of more interest to teens and young adults (demographic clusters 1 and 3). This results in a further confusion of the classifier, and makes a classification based on visual features very difficult.

## Experiment 6 - Demographics estimation using semantic concepts

As Experiment 5 has shown, the direct visual learning of demographic clusters does not lead to strong improvements over random guessing, most likely due to the complex visual nature of the demographic clusters. However, as discussed before, there is a strong connection between semantic concepts and the demographics of their audience. These semantic concepts have a much lower visual complexity than the demographic clusters. Intuitively, separating the demographics estimation from the visual classification could lead to stronger improvements.

To achieve this, the marginalization approach described in Section 5.2 is used. The visual classification is performed by the concept detection systems trained in Experiment 2, whereas a simple demographics model is obtained by measuring

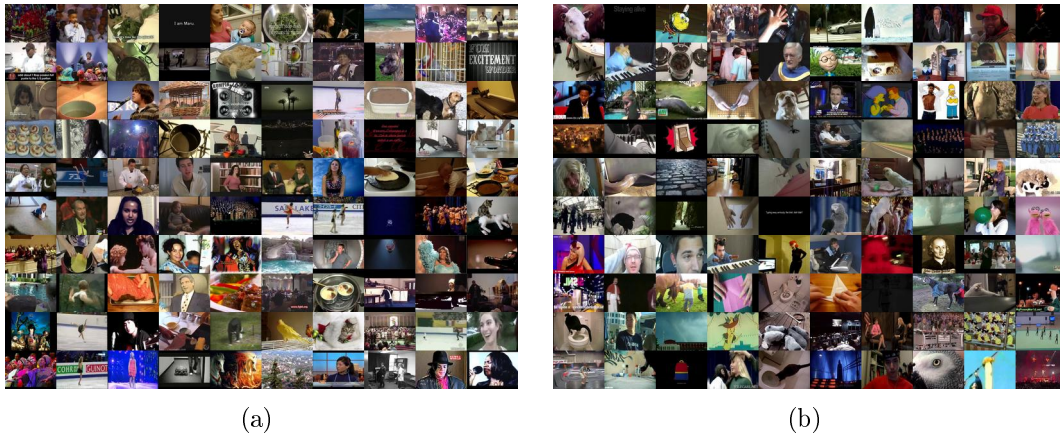


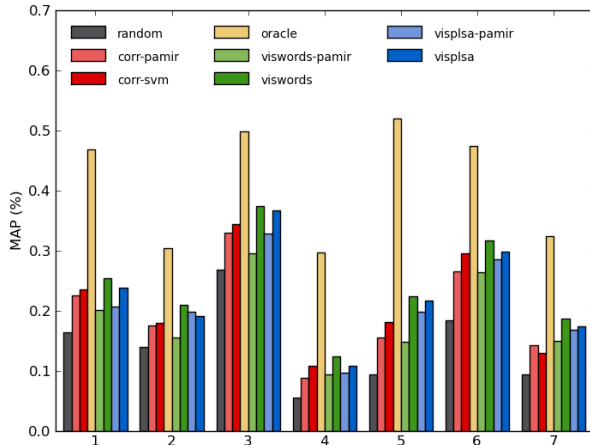
Figure 6.4: Videos from two different demographic clusters, showing a high intra-class diversity.

how each concept is distributed over the demographic clusters. In addition, a control run using a perfect concept detection system is introduced to measure the maximally possible performance of this approach. This 'oracle' system is a theoretical system, which knows the ground truth information of the test set.

The results can be seen in Figure 6.5. First of all, the run with the perfect classifier shows a promising outlook. With a mean average precision of 41.2 %, the estimation of the viewer's age and gender is much more reliable, and could make a lot of videos accessible to demographics-based advertising. All demographic clusters show a 2-fold to 3-fold increase in average precision, but especially the thinner populated clusters 4, 5 (both dominated by females) and 7 (with many users over 35 years old) can be targeted much more accurately with this approach.

When using the best 'real' concept detection system from Experiment 2 (SVM over 2000-dimensional visual words), a mean average precision of 24.1% is achieved. While this is almost a 20 % drop from the oracle system, it still outperforms the baseline system from Experiment 5 by 6.6 %. The drop is notably lower for the densely populated clusters cluster 2 and 3 (mostly focused on male users between 18 and 34), which together hold more than 40 % of all videos. Although the demographics estimation loses a lot of accuracy in this real-world scenario, for some of the clusters the average precision improves more than 10 % over random guessing.

The massive drop in accuracy is can be attributed to the overall low accuracy of the concept detection systems. Many concepts that show a strong orientation towards a specific demographic cluster cannot be detected reliably, like 'dancing' (1.6 % average precision) or 'baby' (2.5 % average precision). It seems plausible



(a)

	SVM	PAMIR
correlograms	21.1	19.8
viswords(2000)	24.1	18.7
viswords(80)	22.8	21.1
random	14.3	
baseline	17.5	
oracle	41.2	

(b)

Figure 6.5: Demographics estimation results (% mean average precision) for the marginalization system. (a) shows the detailed results on cluster level, (b) the overall results.

that with an overall improved concept detection performance the distance to the 'oracle' system can be reduced.

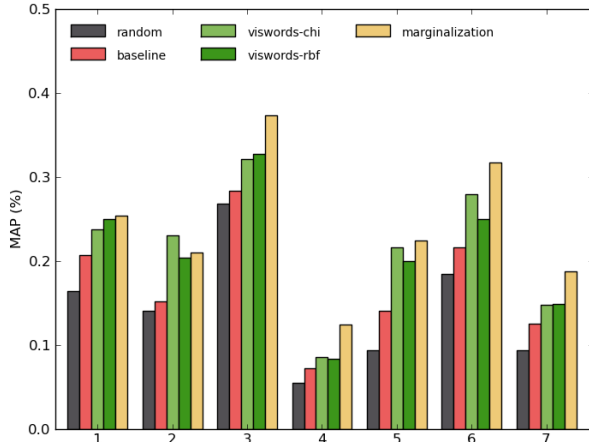
Overall, the mean average precision could be improved by 9.8 % over random guessing (14.3 % MAP). In the following experiment, another approach to obtain a demographics estimate is evaluated, which learns the relation between semantic concepts and demographics using a supervised classifier.

## Experiment 7 - Two-step demographics estimation

In the marginalization approach the estimate was based on concept detection results and a simple demographics model, which was obtained by measuring how the semantic concepts were distributed over the clusters. Another approach, which will be evaluated in this experiment, is to train a second set of classifiers over the concept detection results, learning the relationship between concepts and demographic clusters.

Therefore, the concept detection results from Experiment 2 are used to create a new intermediate 'two-step' feature representation. This is achieved by concatenating the classification scores, as explained in Section 5.2. Because of their superior performance so far, the classification scores are taken from the combination of Support Vector Machine and 2000-dimensional visual words.





(a)

	<i>MAP</i>
$\chi^2$ -SVM	21.7
<i>RBF</i> -SVM	20.9
marginalization	24.1
baseline	17.1
random	14.3

(b)

Figure 6.6: Demographics estimation results (% mean average precision) for the system using the intermediate 'two-step' feature obtained from classification scores. (a) shows the detailed results on cluster level, (b) the overall results.

This 'two-step' feature is then used to learn a Support Vector Machine for each of the demographic clusters. Both the previously applied  $\chi^2$ -kernel and the *RBF*-kernel are used in this experiment.

The results are shown in Figure 6.6, with the best baseline and marginalization-based result for comparison. With a mean average precision of 21.7 % for the  $\chi^2$ -kernel, this systems achieves a 7.4 % improvement over random guessing and a 4.6 % improvement over the baseline approach from Experiment 5. It is, however, outperformed by the marginalization approach from the previous experiment, which was evaluated to achieve a mean average precision of 24.1 %. The use of the *RBF*-kernel does not improve the results.

Again, an issue that was already discussed in Experiment 5 might have a negative impact on the performance of this approach. Many concepts are visually similar, although they attract different demographics. This negatively influences the classifier training, as both negative and positive samples might have a similar 'two-step'-feature representation. Consequently, the overall accuracy is lower.

But while the overall mean average precision of 21.7 % may not seem high, it still gives a solid estimate of the audience's demographics. However, this approach requires a second set of classifiers to be trained, resulting in an increased learning time compared to the other approaches. Given the increased complexity but lower accuracy compared to the marginalization-based approach, it is an unlikely

option for a real-world application.

## Experiment 8 - Concept Vocabulary

The demographics estimation systems proposed in Experiments 6 and 7 are based on the observation that semantic concepts and demographic clusters are closely related. Therefore, the results of a concept detection system were used to obtain the demographics estimate, either by marginalization or by learning a second supervised classifier. Obviously, the quality of the estimate strongly depends on the selection of concepts.

The suitability of a concept for these approaches can be described by different means. First, it is related to the average precision of the corresponding classifier. The higher the achieved average precision, the more likely this concept is to contribute positively to the demographics estimate. This was shown by the 'oracle' run in Experiment 6, where the use of a perfect concept detection system leads to an overall strong improvement compared to the 'real' systems.

Also, how a concept is distributed over the different demographic clusters plays an important role. A concept that attracts a very broad demographics does not help much in narrowing down the potential audience. In conclusion, removing concepts that show a low concept detection accuracy or a wide distribution over the demographic clusters should improve the demographics estimation.

Therefore, the marginalization-based system is now run several times, removing more and more concepts from the vocabulary and observing the change in mean average precision of the demographics estimation. Based on the discussion above, the criteria for a concept to be removed are a low detection accuracy and a wide spread over the demographic clusters. The latter will be measured using the entropy over the cluster distribution for that concept.

The results can be found in Figure 6.7. It can be seen that the removal of a certain number of concepts with low detection accuracy or high entropy (which indicated that they are more widely spread over the demographic clusters) leads to a small improvement for the demographics estimation.

However, a large number of concepts can be removed before the accuracy or the demographics estimation starts to drop below the starting value. Only after removing between 60 and 70 concepts (depending on the criterion), the mean average precision drops rapidly. This implies that many of the dropped concepts could be redundant for the demographics estimation, or simply that a small number of concepts is sufficient enough. A reason for this might be cross-concept effects: a concept detector for a concept like 'interview' also gives high scores to other concepts that predominantly show persons, like 'obama', 'mccain' or 'talkshow'.

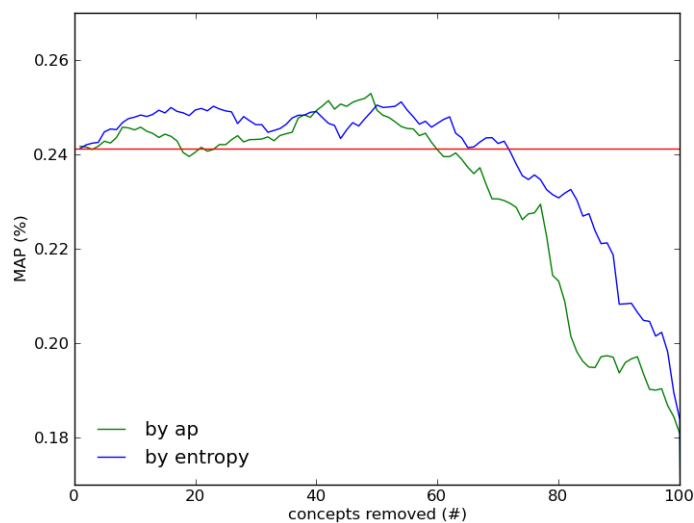


Figure 6.7: Change of mean average precision for the demographics estimate when removing 'bad' concepts. Concepts are removed because of low detection accuracy (green line) or high entropy (blue line). The red line shows the mean average precision with the full set of concepts.

So removing some of these concepts has only a minor impact on the demographics estimate, as 'redundant' detectors can take their place. Surprisingly, with only 5 concepts (namely 'counterstrike-game', 'skateboarding', 'cheerleading', 'horse' and 'mccain') the marginalization approach is still able to outperform the baseline from Experiment 5.



# Chapter 7

## Discussion

The experimental results presented in this thesis have shown that estimating the demographics of a web video's audience by only using visual features presents a difficult challenge.

However, of the three proposed approaches to obtain the estimate, two performed reasonably well. This makes it possible to utilize demographic information in an online advertising system, even if no personal information about the viewer is accessible and without the need for invading their privacy by tracking the browsing behaviour.

With a marginalization-based approach using the results from a Support Vector Machine over visual words, the correct demographic cluster could be estimated with a mean average precision of 24.1 %, a 9.8 % improvement over random guessing. If a very fast processing speed is required, the combination of color correlograms and PAMIR could still achieve a MAP of 19.8 %.

Given the results when using a theoretical, perfect classifier (41.2 % MAP), this method still shows a lot of potential for further improvements. Obviously, a higher mean average precision of the concept detection system would improve the quality of the demographics estimate. This could be achieved by different means, for example by introducing more modalities like audio features [64, 59] or by using a more complex classifier configuration. So further progress in video concept detection has a direct, positive influence on the presented approach. Also, the use of textual information (for example, the video title or keywords, if available) has not been explored in this thesis.

This 'limit' of 41.2 % could be explained through the difficulty of the YouTube dataset. Many semantic concepts are ambiguous, for example, a video showing a piano player cannot only be found within the concept 'piano', but also in other related concepts like 'concert' or 'americas-got-talent' (see Figure 7.1), thus

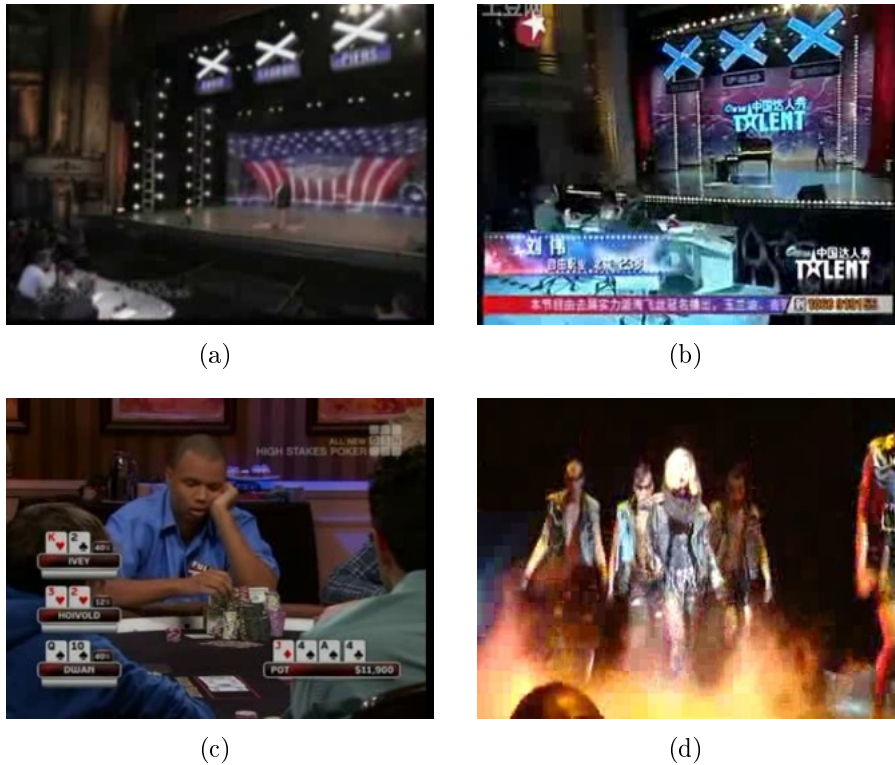


Figure 7.1: Examples that show some of the problems with the dataset: (a) shows a scene from Americas Got Talent and was taken from the corresponding concept, (b) shows a similar scene from the chinese version of the show, but was taken from 'piano'. (c) and (d) were both taken from 'poker', but while one shows the correct activity, the other shows a concert where the song 'Pokerface' was performed.

confusing the concept detectors. Also, the queries that were used to obtain the videos could be improved, to avoid that videos are downloaded that are not related to the concept. For instance, a simple query for 'poker' (as it was used in this thesis) also returns music-videos and concerts where the song 'Pokerface' is performed (see Figure 7.1).

The last experiment has shown that a fixed vocabulary is not the best choice, as concepts that cannot be detected reliably or that attract a broad demographics should be avoided. Both have a negative influence on the demographics estimation. So one question that remains is how a suitable vocabulary can be created. Starting with a very large vocabulary and reducing it according to the aforementioned conditions is one possibility. As an alternative, researchers have recently proposed an automatic discovery of concepts that could also be applied in this scenario [4].

On the other hand, the mean average precision of a state-of-the art concept detection system could be improved from 8.8 % to 12.9 % by using demographic context. This is a remarkable improvement compared to results achieved by fusing multiple visual features (for example, see [42]).

In our case, we were also limited to the available comments on a video. So for many videos the demographic context obtained through the demographic profiles can be considered sparse. With access to detailed viewing data (which is always much larger than the available comments, as discussed before), a further improvement is to be expected. Also, further demographic signals could be obtained by analysing the available textual information, especially the texts of the comments. This was already demonstrated in combination with web blogs [49].

In conclusion, this thesis showed that the visual context on a web video portal can serve to get a solid demographics estimate where not other means of obtaining it are available, and thus becomes an interesting option as additional input for an online advertising system. Also, in cases where demographic context is readily available, it should be used to boost the accuracy of an existing concept detection system.





# Bibliography

- [1] [http://www.youtube.com/results?search\\_query=\\*](http://www.youtube.com/results?search_query=*), 2011.
- [2] [http://www.youtube.com/t/press\\_statistics/](http://www.youtube.com/t/press_statistics/), 2011.
- [3] <http://googleresearch.blogspot.com/2011/06/google-at-cvpr-2011.html>, 2011.
- [4] Hrishikesh Aradhye, George Toderici, and Jay Yagnik. Video2text: Learning to annotate video content. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 144–151, Washington, DC, USA, 2009. IEEE Computer Society.
- [5] M. Baglioni, U. Ferrara, A. Romei, S. Ruggieri, F. Turini, and Via F. Buonarroti. Preprocessing and mining web log data for web personalization. In *8th Italian Conf. on Artificial Intelligence vol. 2829 of LNCS*, pages 237–249. Springer-Verlag, 2003.
- [6] Erwin M. Bakker and Michael S. Lew. Semantic video retrieval using audio analysis. In *Proceedings of the International Conference on Image and Video Retrieval*, CIVR '02, pages 271–277, London, UK, UK, 2002. Springer-Verlag.
- [7] Damian Borth, Adrian Ulges, Markus Koch, and Thomas Breuel. DFKI and university of kaiserslautern participation at TRECVID 2010 - semantic indexing task. In *Proceedings of the TREC Video Retrieval Evaluation Workshop 2010*. NIST, 11 2010. Online-Diskussion mit Mailinglisten.
- [8] Damian Borth, Adrian Ulges, Christian Schulze, and Thomas Breuel. Keyframe extraction for video tagging and summarization. In Gesellschaft fuer Informatik, editor, *Informatiktage 2008*, pages 45–48. GI, 3 2008.
- [9] Theodoros Bozios, Georgios Lekakos, and Victoria Skoularidou. Advanced techniques for personalized advertising in a digital TV environment: The imedia system. In *Proceedings of the eBusiness and eWork Conference*, 2001.

- [10] Darin Brezeale and Diane J. Cook. Learning video preferences from video content. In *Proceedings of the 8th international workshop on Multimedia data mining*, MDM '07, pages 4:1–4:9, New York, NY, USA, 2007. ACM.
- [11] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM.
- [12] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2:121–167, June 1998.
- [13] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [14] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, December 2006.
- [15] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, 1 edition, March 2000.
- [16] Koenw. De Bock and Dirk Van den Poel. Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundam. Inf.*, 98:49–70, January 2010.
- [17] H. Drucker, D. Wu, and V. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5), 1999.
- [18] Shih fu Chang, R. Manmatha, and Tat seng Chua. Combining text and audio-visual features in video indexing. In *In IEEE ICASSP*, 2005.
- [19] Laura Goldberg. Internet ad revenues at nearly \$15 billion in first-half 2011, up 23%, second quarter 2011 breaks record again. [http://www.iab.net/about\\_the\\_iab/recent\\_press\\_releases/press\\_release\\_archive/press\\_release/pr-092811](http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-092811), 2011.
- [20] Avi Goldfarb and Catherine Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30:389–404, May 2011.
- [21] David Grangier and Samy Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1371–1384, August 2008.
- [22] Kristen Grauman and Bastian Leibe. *Visual Object Recognition*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

- [23] Steve R. Gunn. Support vector machines for classification and regression. Technical report, Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, May 1998.
- [24] Alexander Haubold and Apostol Natsev. Web-based information content and its application to concept-based video retrieval. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 437–446, New York, NY, USA, 2008. ACM.
- [25] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.
- [26] Hollis. Ten years of learning on how online advertising builds brands. *Advertising Research*, 45:255–268, 2005.
- [27] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.
- [28] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 151–160, New York, NY, USA, 2007. ACM.
- [29] Jing Huang. *Color-spatial image indexing and applications*. PhD thesis, Ithaca, NY, USA, 1998.
- [30] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 762–, Washington, DC, USA, 1997. IEEE Computer Society.
- [31] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.
- [32] Anil K. Jain. Data clustering: 50 years beyond k-means. In *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, ECML PKDD '08, pages 3–4, Berlin, Heidelberg, 2008. Springer-Verlag.
- [33] Bernard J. Jansen and Tracy Mullen. Sponsored search: an overview of the concept, history, and technology. *IJEB*, 6(2):114–131, 2008.
- [34] Hongjun Jia and A.M. Martinez. Support vector machines in face recognition with occlusions. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:136–141, 2009.

- [35] Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann. Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study. *Multimedia, IEEE Transactions on*, 12(1):42–53, November 2009.
- [36] Moataz Ali Johann Hofmann. An extensive approach to content based image retrieval using low- & high-level descriptors. Master’s thesis.
- [37] Anísio Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani, and Berthier Ribeiro-Neto. Learning to advertise. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’06, pages 549–556, New York, NY, USA, 2006. ACM.
- [38] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV ’99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.
- [39] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.
- [40] Tao Mei, Xian-Sheng Hua, and Shipeng Li. Contextual in-image advertising. In *Proceeding of the 16th ACM international conference on Multimedia*, MM ’08, pages 439–448, New York, NY, USA, 2008. ACM.
- [41] Tao Mei, Xian-Sheng Hua, and Shipeng Li. Videosense: a contextual in-video advertising system. *IEEE Trans. Cir. and Sys. for Video Technol.*, 19:1866–1879, December 2009.
- [42] Chong-Wah Ngo, Yu-Gang Jiang, Xiao-Yong Wei, Feng Wang, Wanlei Zhao, Hung-Khoon Tan, and Xiao Wu. Experimenting vireo-374: Bag-of-visual-words and visual-based ontology for semantic video indexing and search. In Paul Over, George Awad, Wessel Kraaij, and Alan F. Smeaton, editors, *TRECVID*. National Institute of Standards and Technology (NIST), 2007.
- [43] William S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, December 2006.
- [44] Roberto Paredes, Adrian Ulges, and Thomas M. Breuel. Fast discriminative linear models for scalable video tagging. In *ICMLA*, pages 571–576, 2009.
- [45] J. M. Peña, J. A. Lozano, and P. Larrañaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recogn. Lett.*, 20:1027–1040, October 1999.
- [46] Mika Rautiainen, Timo Ojala, and Tapio Seppänen. Analysing the performance of visual, concept and text features in content-based video retrieval.

In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '04, pages 197–204, New York, NY, USA, 2004. ACM.

- [47] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Golgher, and Edleno Silva de Moura. Impedance coupling in content-targeted advertising. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 496–503, New York, NY, USA, 2005. ACM.
- [48] Erin Read Ruddick. How to tailor ads to demographic-based preferences. <http://www.marketingsherpa.com/content/?q=node/5768/>, 2008.
- [49] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of Age and Gender on Blogging. In *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, March 2006.
- [50] Jeremy Scott. 25 jawdropping youtube video facts, figures, statistics. <http://www.reelseo.com/youtube-statistics/#ixzz1ZdNodE00>, 2011.
- [51] Srinivasan H. Sengamedu, Neela Sawant, and Smita Wadhwa. vadeo: video advertising system. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 455–456, New York, NY, USA, 2007. ACM.
- [52] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In Jean Ponce, Martial Hebert, Cordelia Schmid, and Andrew Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 127–144. Springer, 2006.
- [53] Cees G. M. Snoek. Early versus late fusion in semantic video analysis. In *In ACM Multimedia*, pages 399–402, 2005.
- [54] Cees G. M. Snoek and Marcel Worring. Concept-based video retrieval. *Found. Trends Inf. Retr.*, 2:215–322, April 2009.
- [55] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, pages 421–430, New York, NY, USA, 2006. ACM.
- [56] H. Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1:801–804, 1956.
- [57] Adrian Ulges, Markus Koch, Damian Borth, and Thomas M. Breuel. Tubetagger - youtube-based concept detection. In *Proceedings of the 2009 IEEE*

- International Conference on Data Mining Workshops, ICDMW '09*, pages 190–195, Washington, DC, USA, 2009. IEEE Computer Society.
- [58] Chingning Wang, Ping Zhang, Risook Choi, and Michael D’Eredita. Understanding consumers attitude toward advertising. In *In: Eighth Americas Conference on Information Systems. (2002) 1143â1148*, pages 1143–1148, 2002.
- [59] Yao Wang, Zhu Liu, and Jin-Cheng Huang. Multimedia content analysis using both audio and visual cues, 2000.
- [60] Xiaohui Wu, Jun Yan, Ning Liu, Shuicheng Yan, Ying Chen, and Zheng Chen. Probabilistic latent semantic user segmentation for behavioral targeted advertising. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising, ADKDD '09*, pages 10–17, New York, NY, USA, 2009. ACM.
- [61] Xiaoyuan Wu and Alvaro Bolivar. Keyword extraction for contextual advertisement. In *Proceeding of the 17th international conference on World Wide Web, WWW '08*, pages 1195–1196, New York, NY, USA, 2008. ACM.
- [62] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web, WWW '09*, pages 261–270, New York, NY, USA, 2009. ACM.
- [63] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, and Winston Hsu. Columbia university’s baseline detectors for 374 lscm semantic visual concepts. Technical report, Columbia University, March 2007.
- [64] Linjun Yang, Jiemin Liu, Xiaokang Yang, and Xian-Sheng Hua. Multimodality web video categorization. In *Proceedings of the international workshop on Workshop on multimedia information retrieval, MIR '07*, pages 265–274, New York, NY, USA, 2007. ACM.
- [65] Wen-tau Yih, Joshua Goodman, and Vitor R. Carvalho. Finding advertising keywords on web pages. In *Proceedings of the 15th international conference on World Wide Web, WWW '06*, pages 213–222, New York, NY, USA, 2006. ACM.
- [66] James Zern. <http://youtube-global.blogspot.com/2011/04/mmm-mmm-good-youtube-videos-now-served.html>, 2011.
- [67] J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73:2007, 2007.

# Appendix A

## Semantic concept vocabulary

Table A.1: A list of all semantic concepts used and the corresponding YouTube queries. Concepts in bold were used in the experiments, all others dropped because of an insufficient number of comments.

concept	query	category
<b>airplane-flying</b>	airplane & flying -indoor	-
<b>americas-got-talent</b>	americas got talent	-
<b>anime</b>	anime mix	-
<b>aquariums</b>	aquarium fish tank	Animals
arcade	arcade	Travel
<b>asians</b>	asians -hot -sexy -bikini	People
autumn	autumn colors	Travel
<b>baby</b>	baby first	People
badlands	badlands	Travel
<b>balloons</b>	balloons	Entertainm.
<b>baseball</b>	baseball -golf	Sports
<b>basketball</b>	basket ball	Sports
<b>beach</b>	beach	Travel
beehive	beehive	Animals
bicycle	bicycle	Vehicles
<b>bikini</b>	bikini	
<b>bill-clinton</b>	bill clinton	News
<b>birds</b>	birds	Animals
blacksmithing	blacksmith	Howto
boat	boat small -rc	Vehicles
boat-ship	ship &(queen freedom royal)	Vehicles
<b>boobs</b>	boobs tits	
<b>boxing</b>	boxing	Sports
<b>breakdancing</b>	break dancing	
bridge	bridge -crossing -ship	Travel
brown-bear	brown bear	Animals
bus	bus -van -suv -vw -ride	Vehicles
Continued on next page		

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries. Concepts in bold were used in the experiments, all others dropped because of an insufficient number of comments.

concept	query	category
<b>cake</b>	cake	Howto
camels	camel dromedar -spider	Animals
campus	university campus tour	-
<b>car</b>	car	Vehicles
<b>car-crash</b>	car crash	Vehicles
car-racing	car racing -rc	Sports
<b>cartoon</b>	cartoon	Film
castle	castle & (afar outside) -inside	Travel
cathedral	cathedral	Travel
<b>cats</b>	cats	Animals
celebration	celebration	Travel
<b>cheerleading</b>	cheerleading	-
<b>choir</b>	choir	-
<b>christmas-tree</b>	christmas tree -fire	-
<b>circus</b>	circus show	-
city-skyline	skyline	Travel
<b>cityscape</b>	cityscape -slideshow -emakina	Travel
classroom	classroom & school -secret	-
clock-tower	clock tower	Travel-
clouds	clouds & beautiful	Travel
<b>cockpit</b>	cockpit -railway -line	Vehicles
<b>commercial</b>	commercial -barack	-
<b>concert</b>	concert	Music
<b>cooking</b>	cooking	Howto
<b>counterstrike-game</b>	counterstrike movie -lego -real	
<b>court</b>	court judge	News
<b>cows</b>	cow	Animals
crane	crane	Vehicles
<b>crash</b>	crash	Vehicles
dam	dam	Travel
<b>dancing</b>	dancing	People
<b>dark-skinned-people</b>	black people	-
<b>darth-vader</b>	darth vader	-
<b>demonstration</b>	protesting	-
desert	desert	Travel
<b>dog</b>	dog	Animals
<b>dogs</b>	dogs	Animals
<b>drawing</b>	drawing	Film
drinking	drinking competition	-
<b>driver</b>	car & vehicle & driver -simulator	-
drummer	drummer	Howto
eiffeltower	eiffeltower	Travel
emergency-vehicle	emergency & vehicle -driver -ride	Vehicles
excavation	excavation	Travel
Continued on next page		



Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries. Concepts in bold were used in the experiments, all others dropped because of an insufficient number of comments.

concept	query	category
<b>explosion</b>	explosion	Howto
fence	fence	Travel
fencing	fencing	Sports
ferarri	ferarri	Vehicles
firefighter	firefighter training	-
fireworks	fireworks (nice or beautiful)	-
<b>fish</b>	fish	Animals
<b>fishing</b>	fishing	Sports
flood	flood water	News
flower	flower & (bouquet bloom)	-
<b>food</b>	food delicious	-
<b>football</b>	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
<b>georgewbush</b>	george w bush	News
geyser	geyser	Travel
glacier	glacier	Travel
glasses	glasses wearing -not -are	-
<b>golf</b>	golf	Sports
golf-course	golf course flyover	Sports
<b>graffiti</b>	graffiti	-
grand-canyon	grand canyon	Travel
<b>gym</b>	gym	Sports
<b>gymnastics</b>	gymnastics	Sports
hand	hand & daft	-
harbor	harbor & dock	Travel
<b>helicopter</b>	helicopter	Vehicles
highway	highway us route	-
hiking	hiking	Travel
<b>horse</b>	horse	Animals
horse-racing	horse racing	Sports
hospital	hospital & emergency	-
hotel-room	"hotel room"	Travel
house	house sightseeing	Travel
<b>ice-skating</b>	ice skating	Sports
<b>interview</b>	interview	News
<b>iphone</b>	iphone	-
jewellery	jewellery	-
jungle	jungle tropical	Travel
Continued on next page		

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries. Concepts in bold were used in the experiments, all others dropped because of an insufficient number of comments.

concept	query	category
<b>kiss</b>	kissing two	-
<b>kitchen</b>	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	-
market	market	Travel
<b>mccain</b>	john mc cain	News
memorial	memorial -day	Travel
<b>military-parade</b>	military parade	-
<b>monitor</b>	screen monitor	-
<b>moon</b>	moon footage	-
mosque	mosque	Travel
<b>motorcycle</b>	(motorcycle or motorbike) -crash	Vehicles
mountain	mountain & panorama	Travel
<b>muppets</b>	muppet show	-
<b>music-video</b>	music video	-
native-american	native american dance	-
neon-sign	neon sign	Travel
nighttime	"by night"	Travel
<b>obama</b>	barrack obama	News
<b>office</b>	office working	-
<b>old-people</b>	"old people"	-
operating-room	operating room	-
<b>orchestra</b>	orchestra symphony	-
<b>origami</b>	origami	Howto
<b>outer-space</b>	universe galaxy -super -song	-
pagoda	pagoda	Travel
parachute	parachute -no	Sports
<b>penguin</b>	penguin	Animals
<b>phone</b>	phone & device	-
<b>piano</b>	piano playing	-
pier	pier	Travel
playground	playground	Travel
<b>poker</b>	poker	Entertainm.
polar-bear	polar bear	Animals
<b>pope</b>	pope benedict	-
pottery	pottery	-
<b>press-conference</b>	press conference	News
procession	procession	Travel
Continued on next page		

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries. Concepts in bold were used in the experiments, all others dropped because of an insufficient number of comments.

concept	query	category
pyramids	pyramid	Travel
<b>race</b>	race	Vehicles
railroad	railroad train -model	Vehicles
rainbow	rainbow beautiful	Travel
rainforest	rain forest	Travel
ranch	ranch	Travel
<b>rc-car</b>	rc car	Vehicles
restaurant	restaurant	Travel
rice-terrace	rice terrace	Travel
<b>riding</b>	horse riding	-
<b>riot</b>	riot	News
river	river	Travel
<b>robot</b>	robot -dance -dancers	-
<b>rocket-launching</b>	rocket launch -model -mini -toy	-
rodeo	rodeo bull riding	Sports
rooftop	rooftop	Travel
<b>rugby</b>	rugby	Sports
ruins	ruins -underwater	Travel
<b>runway</b>	runway airport	-
safari	safari	Travel
sailing	sailing	Travel
santa	santa (costume or outfit)	-
secondlife	secondlife	Games
shipwreck	ship wreck	Travel
<b>shooting</b>	shooting gun	-
shopping-mall	shopping (mall or center)	Travel
<b>simpsons</b>	the simpsons homer	-
<b>singing</b>	singing & (gospel choire)	-
<b>skateboarding</b>	skateboarding	-
<b>skiing</b>	skiing -water	Sports
sky	beautiful sky	Travel
<b>snake</b>	snake	Animals
snooker	snooker	Sports
<b>soccer</b>	soccer	Sports
<b>soldiers</b>	soldiers -child	News
stairs	stairs	Travel
steppe	steppe	Travel
street	street & paved	-
submarine	submarine	Vehicles
subway	subway station	Travel
sunrise	sunrise	Travel
<b>surfing</b>	surfing wave	-
swimming	swimming	Sports
swimming-pools	swimming pool	Travel
Continued on next page		

Table A.1: (Continued) A list of all semantic concepts used and the corresponding YouTube queries. Concepts in bold were used in the experiments, all others dropped because of an insufficient number of comments.

concept	query	category
sword-fight	sword fight	Sports
<b>talkshow</b>	talkshow	People
<b>tank</b>	tank	Vehicles
<b>tennis</b>	tennis -table	Sports
tent	tent	Travel
themepark	park &(amusement theme)	Travel
<b>toilet</b>	toilet	-
tony-blair	tony blair	News
<b>tornado</b>	tornado	-
tractor-combine	(harvester or tractor)	Vehicles
traffic	traffic	Travel
traffic-lights	traffic lights	Travel
tunnel	tunnel+ &(through inside)	Travel
	-approach	
turban	turban	-
<b>two-people</b>	two & people -sleepy -questions	-
underwater	underwater	Travel
us-flag	US flag raised	-
vending-machine	vending machine	Travel
<b>videoblog</b>	videoblog	People
waterfall	waterfall	Travel
weather	weather forecast	-
wedding	wedding footage	-
<b>wheel</b>	wheel	Vehicles
windmill	wind mill	Travel
<b>windows-desktop</b>	windows desktop	-
<b>worldofwarcraft</b>	world of warcraft	Entertainm.
<b>wrestling</b>	wrestling	Sports

## Detailed concept detection results

Table A.2: Detailed results for all evaluated experiments for visual features. The experiments were performed on the final set of 105 semantic concepts.

	SVM			PAMIR		
	color corr.	viswords (2000)	viswords (80)	color corr.	viswords (2000)	viswords (80)
airplane-flying	0.028	0.035	<b>0.038</b>	0.026	0.026	0.031
americas-got-talent	<b>0.300</b>	0.222	0.194	0.203	0.119	0.040
anime	0.120	<b>0.182</b>	0.111	0.046	0.111	0.081
aquariums	0.062	<b>0.167</b>	0.113	0.101	0.111	0.065
asians	0.035	<b>0.066</b>	0.042	0.018	0.049	0.041
baby	0.017	<b>0.025</b>	0.021	0.014	0.024	0.021
balloons	0.013	<b>0.019</b>	0.014	0.015	0.013	0.015
baseball	<b>0.033</b>	0.033	0.019	0.018	0.024	0.017
basketball	0.072	<b>0.084</b>	0.077	0.035	0.068	0.035
beach	0.039	<b>0.041</b>	0.031	0.031	0.025	0.024
bikini	0.014	<b>0.023</b>	0.020	0.013	0.018	0.018
bill-clinton	0.040	<b>0.126</b>	0.047	0.035	0.063	0.046
birds	0.020	<b>0.040</b>	0.025	0.023	0.019	0.023
boobs	0.012	<b>0.020</b>	0.019	0.013	0.019	0.019
boxing	0.102	<b>0.321</b>	0.233	0.040	0.247	0.124
breakdancing	0.096	<b>0.169</b>	0.108	0.019	0.113	0.059
cake	0.031	<b>0.081</b>	0.055	0.025	0.061	0.049
car	0.018	<b>0.049</b>	0.030	0.020	0.032	0.033
car-crash	0.021	<b>0.047</b>	0.027	0.016	0.034	0.026
cartoon	0.049	<b>0.101</b>	0.048	0.031	0.077	0.044
cats	0.026	0.025	0.022	<b>0.028</b>	0.018	0.013
cheerleading	0.038	<b>0.101</b>	0.059	0.022	0.040	0.040
choir	0.029	<b>0.047</b>	0.042	0.030	0.037	0.034
christmas-tree	0.010	<b>0.025</b>	0.016	0.012	0.011	0.017
circus	0.024	<b>0.028</b>	0.022	0.022	0.027	0.028
cityscape	0.029	<b>0.047</b>	0.028	0.021	0.038	0.026
cockpit	0.059	<b>0.174</b>	0.112	0.033	0.112	0.065
commercial	<b>0.020</b>	0.018	0.015	0.015	0.014	0.015
concert	<b>0.089</b>	0.081	0.066	0.051	0.042	0.055
cooking	0.027	<b>0.049</b>	0.038	0.023	0.041	0.033
counterstrike-game	0.082	<b>0.121</b>	0.070	0.043	0.045	0.034
court	0.016	0.024	0.023	0.015	<b>0.031</b>	0.027
cows	0.018	<b>0.044</b>	0.030	0.018	0.035	0.029
crash	0.024	<b>0.040</b>	0.033	0.026	0.037	0.030
dancing	<b>0.016</b>	0.016	0.014	0.016	0.015	0.015
dark-skinned-people	0.010	0.013	<b>0.013</b>	0.010	0.012	0.012
darth-vader	0.130	<b>0.200</b>	0.151	0.038	0.113	0.065
demonstration	0.030	0.036	<b>0.037</b>	0.018	0.030	0.025
dogs	0.019	<b>0.022</b>	0.012	0.018	0.021	0.016

Continued on next page

Table A.2: (Continued) Detailed results for all evaluated experiments for visual features. The experiments were performed on the final set of 105 semantic concepts.

	SVM			PAMIR		
	color corr.	viswords (2000)	viswords (80)	color corr.	viswords (2000)	viswords (80)
drawing	0.073	<b>0.199</b>	0.127	0.055	0.113	0.086
driver	0.014	<b>0.034</b>	0.026	0.015	0.023	0.023
explosion	<b>0.066</b>	0.047	0.035	0.013	0.020	0.021
fish	0.019	0.018	0.017	<b>0.019</b>	0.018	0.017
fishing	0.026	<b>0.089</b>	0.041	0.025	0.040	0.034
food	0.027	0.024	0.017	<b>0.027</b>	0.023	0.018
football	0.086	<b>0.120</b>	0.082	0.065	0.062	0.053
georgewbush	0.033	<b>0.083</b>	0.059	0.027	0.044	0.043
golf	<b>0.086</b>	0.059	0.047	0.069	0.032	0.026
graffiti	0.016	0.021	0.017	0.013	<b>0.022</b>	0.018
gym	0.012	<b>0.042</b>	0.026	0.016	0.037	0.027
gymnastics	0.036	<b>0.073</b>	0.050	0.032	0.034	0.035
helicopter	0.039	<b>0.076</b>	0.065	0.033	0.042	0.036
horse	0.027	<b>0.070</b>	0.046	0.026	0.038	0.030
ice-skating	0.140	<b>0.338</b>	0.281	0.051	0.244	0.170
interview	0.034	<b>0.047</b>	0.042	0.024	0.038	0.028
iphone	0.045	<b>0.084</b>	0.052	0.029	0.051	0.028
kiss	<b>0.024</b>	0.020	0.020	0.018	0.019	0.020
kitchen	0.021	0.028	0.028	<b>0.030</b>	0.027	0.027
mccain	0.060	<b>0.110</b>	0.076	0.038	0.069	0.050
military-parade	0.061	<b>0.237</b>	0.151	0.024	0.175	0.081
monitor	0.013	<b>0.061</b>	0.056	0.016	0.052	0.035
moon	0.030	<b>0.047</b>	0.023	0.028	0.017	0.021
motorcycle	0.019	<b>0.073</b>	0.030	0.025	0.048	0.032
muppets	0.146	<b>0.286</b>	0.165	0.105	0.112	0.052
music-video	0.019	<b>0.020</b>	0.018	0.015	0.018	0.017
obama	0.028	0.078	<b>0.085</b>	0.022	0.050	0.033
office	0.015	<b>0.022</b>	0.019	0.011	0.019	0.017
old-people	<b>0.030</b>	0.028	0.026	0.015	0.013	0.014
orchestra	<b>0.184</b>	0.166	0.133	0.103	0.127	0.085
origami	0.048	<b>0.320</b>	0.222	0.112	0.203	0.141
outer-space	0.054	<b>0.197</b>	0.135	0.036	0.080	0.058
penguin	0.023	<b>0.028</b>	0.020	0.016	0.018	0.013
phone	0.049	<b>0.089</b>	0.053	0.029	0.061	0.029
piano	0.031	<b>0.059</b>	0.031	0.017	0.028	0.022
poker	0.102	<b>0.183</b>	0.115	0.064	0.100	0.060
pope	0.019	<b>0.051</b>	0.035	0.030	0.022	0.017
press-conference	0.025	<b>0.068</b>	0.046	0.027	0.053	0.042
race	0.022	0.059	0.058	0.023	<b>0.074</b>	0.055
rc-car	0.015	<b>0.046</b>	0.038	0.015	0.036	0.024
riding	0.043	<b>0.090</b>	0.049	0.026	0.033	0.030
riot	0.024	<b>0.048</b>	0.033	0.017	0.037	0.030
robot	0.013	<b>0.019</b>	0.013	0.012	0.014	0.012

Continued on next page

Table A.2: (Continued) Detailed results for all evaluated experiments for visual features. The experiments were performed on the final set of 105 semantic concepts.

	SVM			PAMIR		
	color corr.	viswords (2000)	viswords (80)	color corr.	viswords (2000)	viswords (80)
rocket-launching	0.040	<b>0.092</b>	0.081	0.029	0.038	0.039
rugby	0.163	<b>0.218</b>	0.158	0.104	0.086	0.084
runway	0.065	<b>0.129</b>	0.096	0.051	0.070	0.093
shooting	0.013	<b>0.027</b>	0.020	0.017	0.020	0.016
simpsons	<b>0.273</b>	0.236	0.191	0.149	0.048	0.025
singing	0.013	<b>0.041</b>	0.028	0.018	0.038	0.031
skateboarding	0.025	<b>0.072</b>	0.047	0.021	0.043	0.030
skiing	0.118	<b>0.149</b>	0.121	0.102	0.048	0.071
snake	0.025	<b>0.043</b>	0.038	0.027	0.026	0.024
soccer	0.085	<b>0.096</b>	0.070	0.053	0.060	0.043
soldiers	0.022	<b>0.026</b>	0.018	0.014	0.015	0.017
surfing	0.118	<b>0.199</b>	0.154	0.061	0.083	0.090
talkshow	0.150	<b>0.345</b>	0.229	0.067	0.303	0.132
tank	0.015	<b>0.057</b>	0.037	0.017	0.039	0.023
tennis	0.060	<b>0.215</b>	0.144	0.031	0.144	0.050
toilet	0.026	<b>0.038</b>	0.035	0.015	0.031	0.022
tornado	0.101	<b>0.157</b>	0.120	0.027	0.047	0.061
two-people	0.012	0.013	0.013	<b>0.014</b>	0.012	0.012
videoblog	0.015	0.025	<b>0.027</b>	0.013	0.023	0.027
wheel	0.015	0.019	0.022	0.014	<b>0.026</b>	0.019
windows-desktop	0.070	0.126	0.103	0.039	<b>0.130</b>	0.081
worldofwarcraft	<b>0.128</b>	0.054	0.038	0.060	0.043	0.026
wrestling	0.023	<b>0.041</b>	0.034	0.022	0.022	0.021
total	0.050	<b>0.088</b>	0.063	0.034	0.055	0.039

## Detailed feature fusion results.

Table A.3: Detailed results for all evaluated experiments for the fusion of demographic information and visual features. The experiments were performed on the final set of 105 semantic concepts.

	early-conc	early-mult	late-avg		
	viswords (2000)	viswords (2000)	color corr.	viswords (2000)	viswords (80)
airplane-flying	0.081	0.076	0.068	<b>0.086</b>	0.084
americas-got-talent	0.126	0.167	<b>0.348</b>	0.228	0.225
anime	0.306	0.326	0.520	<b>0.538</b>	0.440
aquariums	0.106	0.107	0.144	<b>0.212</b>	0.172
asians	<b>0.165</b>	0.127	0.061	0.093	0.078
baby	<b>0.116</b>	0.095	0.092	0.101	0.101
balloons	0.024	0.024	0.045	<b>0.045</b>	0.039
baseball	0.036	0.028	0.034	<b>0.036</b>	0.022
basketball	0.095	<b>0.103</b>	0.094	0.097	0.084
beach	0.085	<b>0.088</b>	0.048	0.072	0.055
bikini	<b>0.027</b>	0.020	0.015	0.025	0.021
bill-clinton	0.124	0.080	0.083	<b>0.129</b>	0.056
birds	0.056	0.048	0.038	<b>0.064</b>	0.043
boobs	<b>0.026</b>	0.026	0.014	0.021	0.021
boxing	0.345	0.299	0.283	<b>0.428</b>	0.355
breakdancing	<b>0.194</b>	0.161	0.077	0.140	0.095
cake	0.274	0.229	0.216	<b>0.332</b>	0.319
car	<b>0.052</b>	0.050	0.024	0.050	0.036
car-crash	0.048	0.038	0.026	<b>0.050</b>	0.032
cartoon	<b>0.125</b>	0.082	0.050	0.096	0.052
cats	0.073	0.065	0.076	0.071	<b>0.082</b>
cheerleading	<b>0.330</b>	0.218	0.150	0.293	0.244
choir	0.119	0.105	0.081	<b>0.125</b>	0.122
christmas-tree	0.013	0.014	0.015	<b>0.029</b>	0.028
circus	<b>0.055</b>	0.033	0.025	0.029	0.022
cityscape	0.044	0.032	0.032	<b>0.051</b>	0.030
cockpit	0.273	0.230	0.165	<b>0.302</b>	0.262
commercial	0.032	<b>0.036</b>	0.036	0.031	0.028
concert	0.063	0.062	0.123	<b>0.126</b>	0.106
cooking	0.073	0.086	0.058	0.089	<b>0.090</b>
counterstrike-game	0.352	0.362	0.482	<b>0.494</b>	0.429
court	<b>0.131</b>	0.098	0.057	0.071	0.067
cows	0.056	<b>0.062</b>	0.021	0.045	0.035
crash	<b>0.082</b>	0.066	0.041	0.063	0.053
dancing	<b>0.042</b>	0.034	0.024	0.032	0.027
dark-skinned-people	<b>0.031</b>	0.015	0.020	0.027	0.028
darth-vader	0.136	0.139	0.119	<b>0.166</b>	0.139
demonstration	<b>0.063</b>	0.038	0.036	0.042	0.039
dogs	<b>0.122</b>	0.093	0.057	0.068	0.064

Continued on next page



Table A.3: (Continued) Detailed results for all evaluated experiments for the fusion of demographic information and visual features. The experiments were performed on the final set of 105 semantic concepts.

	early-conc	early-mult	late-avg		
	viswords (80)	viswords (80)	color corr.	viswords (2000)	viswords (80)
drawing	0.121	0.159	0.084	<b>0.172</b>	0.132
driver	0.029	0.032	0.017	<b>0.039</b>	0.032
explosion	<b>0.102</b>	0.068	0.042	0.081	0.074
fish	<b>0.027</b>	0.017	0.023	0.021	0.020
fishing	0.084	0.086	0.079	<b>0.142</b>	0.088
food	0.038	0.021	0.050	<b>0.050</b>	0.037
football	0.126	0.113	0.084	<b>0.128</b>	0.090
georgewbush	0.035	0.039	0.060	<b>0.078</b>	0.067
golf	0.083	0.047	<b>0.155</b>	0.110	0.084
graffiti	<b>0.075</b>	0.069	0.017	0.023	0.020
gym	<b>0.093</b>	0.074	0.031	0.062	0.053
gymnastics	0.143	0.115	0.149	<b>0.183</b>	0.159
helicopter	<b>0.108</b>	0.097	0.067	0.104	0.089
horse	0.224	0.210	0.227	<b>0.277</b>	0.239
ice-skating	<b>0.373</b>	0.254	0.235	0.363	0.340
interview	<b>0.053</b>	0.045	0.039	0.049	0.043
iphone	<b>0.080</b>	0.051	0.052	0.075	0.054
kiss	0.076	0.064	<b>0.084</b>	0.044	0.041
kitchen	<b>0.081</b>	0.065	0.034	0.041	0.042
mccain	0.140	0.138	0.114	<b>0.162</b>	0.132
military-parade	0.228	0.152	0.123	<b>0.309</b>	0.232
monitor	<b>0.097</b>	0.092	0.013	0.061	0.056
moon	<b>0.059</b>	0.046	0.039	0.041	0.030
motorcycle	0.067	<b>0.083</b>	0.038	0.077	0.050
muppets	0.283	0.244	0.313	<b>0.373</b>	0.267
music-video	<b>0.072</b>	0.044	0.037	0.046	0.042
obama	<b>0.106</b>	0.081	0.060	0.079	0.069
office	0.017	0.016	0.013	<b>0.020</b>	0.018
old-people	<b>0.049</b>	0.032	0.021	0.024	0.026
orchestra	0.186	0.146	<b>0.270</b>	0.239	0.214
origami	0.236	0.263	0.106	<b>0.317</b>	0.269
outer-space	0.167	0.181	0.067	<b>0.191</b>	0.148
penguin	0.061	0.023	0.062	<b>0.081</b>	0.061
phone	0.049	0.069	0.047	<b>0.094</b>	0.054
piano	<b>0.073</b>	0.052	0.030	0.043	0.033
poker	0.344	0.336	0.306	<b>0.376</b>	0.346
pope	0.045	0.045	0.067	<b>0.076</b>	0.068
press-conference	0.132	0.133	0.084	<b>0.146</b>	0.120
race	0.083	<b>0.088</b>	0.030	0.087	0.084
rc-car	0.084	0.078	0.073	<b>0.149</b>	0.140
riding	0.186	0.200	<b>0.313</b>	0.308	0.303
riot	<b>0.089</b>	0.079	0.031	0.052	0.039
robot	0.023	<b>0.030</b>	0.015	0.018	0.017

Continued on next page

Table A.3: (Continued) Detailed results for all evaluated experiments for the fusion of demographic information and visual features. The experiments were performed on the final set of 105 semantic concepts.

	early-conc	early-mult	late-avg		
	viswords (80)	viswords (80)	color corr.	viswords (2000)	viswords (80)
rocket-launching	0.122	0.118	0.060	<b>0.123</b>	0.107
rugby	0.183	0.185	0.145	<b>0.203</b>	0.148
runway	0.193	0.170	0.194	<b>0.256</b>	0.228
shooting	0.036	<b>0.043</b>	0.029	0.035	0.030
simpsons	0.151	0.151	<b>0.282</b>	0.178	0.161
singing	<b>0.191</b>	0.150	0.063	0.189	0.144
skateboarding	<b>0.142</b>	0.115	0.087	0.133	0.118
skiing	<b>0.165</b>	0.099	0.146	0.160	0.135
snake	<b>0.073</b>	0.062	0.034	0.069	0.060
soccer	0.124	0.112	0.137	<b>0.151</b>	0.120
soldiers	<b>0.034</b>	0.030	0.028	0.029	0.023
surfing	0.103	0.125	0.141	<b>0.244</b>	0.204
talkshow	0.332	0.334	0.242	<b>0.348</b>	0.306
tank	<b>0.113</b>	0.096	0.031	0.112	0.084
tennis	<b>0.176</b>	0.149	0.049	0.151	0.097
toilet	<b>0.055</b>	0.041	0.026	0.029	0.037
tornado	0.310	0.269	0.219	<b>0.331</b>	0.265
two-people	0.014	0.014	0.014	<b>0.015</b>	0.014
videoblog	0.035	<b>0.044</b>	0.035	0.041	0.041
wheel	0.039	<b>0.052</b>	0.028	0.048	0.045
windows-desktop	<b>0.127</b>	0.124	0.082	0.123	0.102
worldofwarcraft	0.121	0.088	<b>0.202</b>	0.145	0.104
wrestling	<b>0.050</b>	0.042	0.033	0.048	0.041
total	0.116	0.103	0.095	<b>0.129</b>	0.108

## Detailed demographics estimation results.

Table A.4: Detailed results for all evaluated runs of the baseline system.

	SVM			PAMIR		
	color corr.	viswords (2000)	viswords (80)	color corr.	viswords (2000)	viswords (80)
cluster 1	0.165	0.207	0.203	0.192	<b>0.210</b>	0.198
cluster 2	0.141	<b>0.152</b>	0.149	0.137	0.143	0.151
cluster 3	0.289	0.283	<b>0.295</b>	0.287	0.283	0.288
cluster 4	0.063	0.072	<b>0.077</b>	0.064	0.071	0.069
cluster 5	0.096	0.141	<b>0.142</b>	0.116	0.142	0.141
cluster 6	0.218	0.216	0.221	0.215	0.220	<b>0.225</b>
cluster 7	0.108	0.125	<b>0.135</b>	0.103	0.121	0.122
total	0.154	0.171	<b>0.175</b>	0.159	0.170	0.171

Table A.5: Detailed results for all evaluated runs of the marginalization-based approach system.

	SVM			PAMIR			oracle
	color corr.	viswords (2000)	viswords (80)	color corr.	viswords (2000)	viswords (80)	
cluster 1	0.235	0.254	0.238	0.225	0.201	0.207	0.468
cluster 2	0.180	0.210	0.191	0.176	0.155	0.198	0.304
cluster 3	0.344	0.373	0.367	0.329	0.295	0.328	0.498
cluster 4	0.109	0.124	0.108	0.088	0.094	0.097	0.296
cluster 5	0.181	0.224	0.217	0.156	0.148	0.198	0.520
cluster 6	0.295	0.317	0.299	0.266	0.264	0.285	0.474
cluster 7	0.130	0.187	0.174	0.142	0.150	0.168	0.323
total	0.211	0.241	0.228	0.198	0.187	0.211	0.412

Table A.6: Detailed results for all evaluated runs of the two-step system.

	<i>RBF</i> -SVM	$\chi^2$ -SVM
cluster 1	<b>0.250</b>	0.238
cluster 2	0.204	<b>0.230</b>
cluster 3	<b>0.328</b>	0.321
cluster 4	0.083	<b>0.085</b>
cluster 5	0.199	<b>0.216</b>
cluster 6	0.249	<b>0.280</b>
cluster 7	<b>0.149</b>	0.147
total	0.209	<b>0.217</b>