Int J Multimed Info Retr
DOI 10.1007/s13735-012-0029-x

REGULAR PAPER

Content analysis meets viewers: linking concept detection with demographics on YouTube

Adrian Ulges · Damian Borth · Markus Koch

Received: 2 November 2012 / Accepted: 27 November 2012

Abstract Social image and video sharing provides the opportunity for a user-centric, behavioral auto-understanding of image and video content. We add demographic aspects to this puzzle, i.e. the popularity of content across different ages and genders: employing user comments, we calculate demographic viewership profiles for YouTube clips and provide evidence that these profiles are strongly correlated with semantic concepts appearing in a video. Based on this fact, we outline two approaches that combine video content analysis with demographic aspects: first, we show that concept detection can be used to establish a mapping from content via concepts to viewer demographics (which we refer to as *content-based demographics prediction*). Second, in case sufficient view statistics already give an estimate of a clip's audience, they can be used as a *demographic signal* to disambiguate concept detection in cases of visually similar concepts. We validate the above statements on a dataset of 14,000 YouTube clips covering 105 concepts and commented by 1 mio. users: content-based demographics prediction is shown to provide an accuracy comparable to other information sources (such as a video's tags or uploader data). Also, demographic signals can improve the accuracy of concept detection significantly (by 47 % compared to a content-only approach).

A. Ulges (✉)
German Research Center for Artificial Intelligence (DFKI),
Kaiserslautern, Germany
e-mail: adrian.ulges@dfki.de

D. Borth
University of Kaiserslautern and DFKI, Kaiserslautern, Germany
e-mail: damian.borth@dfki.de

M. Koch
Insiders Technologies, Kaiserslautern, Germany
e-mail: m.koch@insiders-technologies.de

Keywords Concept detection · Video analysis · Demographics prediction · Online video · Content-based video retrieval

1 Introduction

Over the past years, social media sharing has experienced a break-through that fundamentally changes the way we are informed, entertained, and in which we communicate with friends and share personal events. Socially shared content is of tremendous scale and diversity, ranging from personal photos to professionally produced video and across virtually any possible genre. Likewise, the diversity of viewers is enormous: people across all demographic groups come to sites like YouTube with various intentions such as to be entertained, to be educated, to be informed, or to communicate with friends [35].

This development poses both opportunities and challenges to multimedia understanding: though content analysis remains an important topic—after all, the majority of content is sparsely labeled and viewed—new signals of user interaction have entered the scene, be it textual meta-data such as tags and titles, user profiles, comments, ratings, etc. This allows multimedia analysis systems a more holistic understanding of images and video in combination with the intentions and behavior of their audience.

The work presented in this paper follows this line of thought in a sense that we analyze video content in combination with behavioral signals. More precisely, we study the demographic aspects of social video: we present a system that automatically predicts the semantic concepts appearing in a video, together with the popularity of a clip across different user ages and genders. The core component forms a content-based *concept detection* engine [33] that automatically mines web clips for objects, locations and actions appearing in

Published online: 29 December 2012

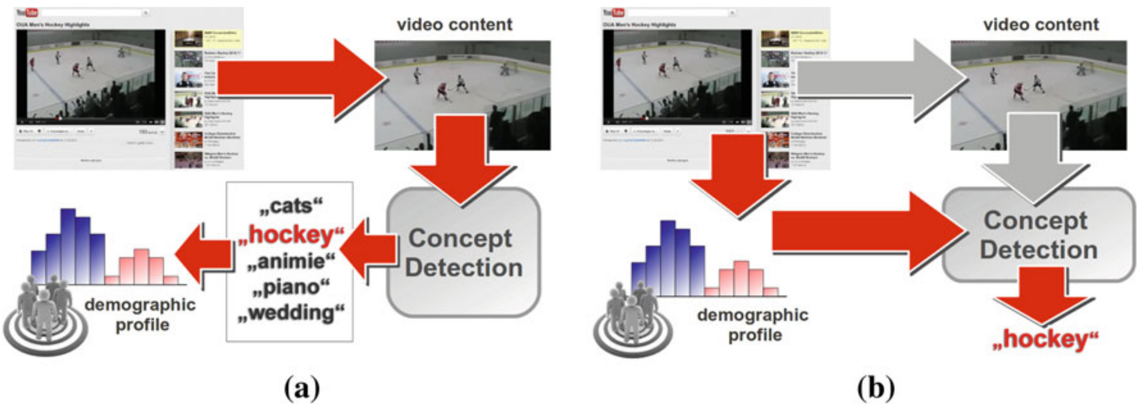

Fig. 1 We study two strategies for linking viewer demographics with concept detection—**a** *Content-based Demographics Prediction*: We auto-detect semantic concepts appearing in the video. As these are often linked to certain demographic user groups, the demographic profile of a video (i.e. its popularity across different ages and genders) can be estimated. **b** *Demographic Signals for Concept Detection*: For other videos—where sufficient view statistics allow the estimation of a demographic profile—this profile can be used as an additional signal to disambiguate concept detection

them. Our key contribution is that we combine this concept detection with demographic information: exploiting user comments on YouTube, we estimate the *demographic viewership profiles* of YouTube clips. We show that these profiles are strongly related to the semantic concepts appearing in the video—for example, the concept "skateboarding" is predominantly viewed by male users, while the concept "cheerleading" is more popular among female ones.

This correlation allows us to link viewership demographics on the one hand with content analysis on the other. We present two strategies that investigate both directions of this link, which will in the following be referred to as *content-based demographics prediction* and *demographic signals for concept detection*.

1.1 Content-based demographics prediction

The first key question we address is *Can we predict from a clip's content which demographic audience it will attract?* This is a vital issue to web video services: video hosting is expensive (in 2009, YouTube's expenses for bandwidth, data center costs, content acquisition, etc. have been quantified to over 2 mio. USD per day [26]), and to cover these costs, web videos are linked with a variety of ads. *Targeting* those ads, i.e. their mapping to certain users or content, is often done via a demographic profiling, based on the assumption that users of different ages and genders show a specific interest in certain products [2,8].

A key challenge to targeting is that the information current approaches ground on is often sparse, such that a large number of clips is missed by targeting (for example, YouTube has been estimated to monetize only 14 % of its views [41]). Here, an automatic content-based demographics prediction—even if limited in accuracy—can be a valuable complement to other information sources, as it is available even in cases where a video lacks a title or tags, where no viewing data are available (if the video has just recently been uploaded) and neither is a user profile of the viewer or uploader (if not logged in, or if no age and gender were specified).

We present an approach that applies *concept detection* and—based on its detection results—infers the popularity of a clip across different viewer groups (an illustration is given in Fig. 1a). Thereby, to take the uncertainty of both concept detection and concept-to-demographics assignment into account, a probabilistic setup is chosen, including a marginalization over latent variables modeling concept presence.

We evaluate the approach on a dataset of 14,000 clips from YouTube (1,300 h of video) commented by about 1 mio. users. Our results indicate that concept detection is a suitable approach when it comes to exploiting video content for demographics estimation: the approach outperforms a direct visual classification of demographic categories as a baseline and performs comparably to other features based on textual meta-data and uploader demographics.

1.2 Demographic signals for concept detection

Instead of inferring demographics, we can also use it as an additional input for content analysis (see Fig. 1 for an illustration): Conventionally, concept detection focuses on the video stream itself as an input, which is bound to fail in case of visually similar concepts (think of "ice hockey" vs. "figure skating"). We suggest using demographic information—i.e. the distribution of popularity across different ages and genders—as an additional feature complementing conventional content-based descriptors. Concept detection can use this feature to improve accuracy: for example, if a video attracts a predominantly male audience, it is much more likely to show the concept "ice hockey". This approach is particularly interesting for videos that have already been viewed and commented on sufficiently. Here, concept detection may be of interest to improve meta-data or to localize concepts at scene level.

Our approach towards including demographic signals into concept detection employs well-established feature fusion mechanisms—note that our contribution lies not in an innovative combination but rather in the investigation of demographics as a novel feature per se. Our experimental results indicate that concept detection can indeed benefit significantly from this novel feature: compared with a content-only approach, a relative improvements of 47 % in accuracy was measured.

This paper is organized as follows: Related work on concept detection and content-based advertising for videos is discussed in Sect. 2. We then describe our approach for estimating demographic user groups and demographic profiles of videos in Sect. 3. After this, two sections introduce the two key contributions of the paper, namely content-based demographics prediction (Sect. 4) and demographic signals for concept detection (Sect. 5). Experimental results for both approaches are outlined in Sects. 6, and 7 concludes the paper.

2 Related work

In the following, research related to our work will be outlined, including concept detection, ad targeting (particularly demographics estimation), as well image and video analysis for advertising.

2.1 Video concept detection

The challenge of automatically detecting semantic concepts such as objects, locations, and activities in video streams—referred to as *video annotation* [9], *concept detection* [39], or *high-level feature extraction* [28]—has been subject to extensive study over the past decade. In benchmarks like TRECVID [28] or the PASCAL visual object challenge [6], the research community has investigated a variety of features and statistical models—please refer to [30] for a survey.

Originally, research in the field has focused on expert-defined tag vocabularies and training data, which are, however, limited in scalability and flexibility. More recent approaches have therefore turned towards portals like Flickr and YouTube as information sources for visual learning, employing user-generated tags as an alternative to expert training labels [14,33,42]. Key research issues include the adaptation to weak label information [16] and the automatic selection of tag vocabularies [9]. The work presented here aligns with this line of research in a sense that web-based tags and content are employed. Our focus, however, is less on concept detection itself but rather on its combination with viewer demographics.

From a technical perspective, our work bears similarities to recent approaches that use concept detection as an intermediate step to model further layers of abstraction: concept detection scores form a *concept signature*, from which *multimedia events* [5,22], or even new classes of attributes [32] are predicted. We follow a similar strategy to infer viewer demographics from concept detection results.

2.2 Ad targeting strategies

A variety of strategies have been proposed for ad targeting, i.e. selection of ads with respect to their potential audience [12,38,40]. Particularly, online advertising has become an indispensable tool for the internet's economic system, and the problem of selecting the most appropriate ad for the visitor of a web page is a topic of extensive research. Of the different approaches studied, *keyword advertising* [12] is the oldest and most popular. It matches keywords given by the user through a search query to a dictionary of sponsored terms—a strategy that has proven highly effective and is implemented in several practical systems (e.g., Google Adwords). As an alternative, *contextual advertising* [36,40] aims at providing even more relevant ads by analyzing the content of a web page and matching the ads accordingly. It removes the necessity of a user input, as it relies on matching ads to the content of a web page. Contextual strategies mostly rely on textual information, although in recent times some research has been presented using image and video content (see below). A third strategy, *behavioral targeting* [3,37], observes a users' web search and browsing behavior by tracking him through user profiles and third party cookies.

The key strategy for successful advertising lies in an optimal combination of the above strategies: In a web video setting, we can combine keyword advertising (e.g., sponsored search on the main page) and behavioral targeting (by profiling users via their view events) with contextual advertising (for example, employing video's tags). The content-based approaches presented in the following can serve as a complement to these strategies, which works even in cases of unknown user profile and weakly labeled videos.

2.3 Demographics estimation

The automatic prediction of users' demographic attributes can be considered the standard when it comes to advertising in "traditional" media (with TV broadcasting being the most prominent example). For example, the revenue of a TV commercial strongly depends on demographic attributes of the attracted audience [4]. Statistics about the demographics of the audience (age, gender, economic class, area), however, are still mostly acquired by monitoring a small subset of the viewers ("Nielsen ratings").

More recently it has been proposed to segment audiences based on interactions on social media pages like Facebook or

Twitter. Demographics estimation has also been studied for web browsing, mainly by training supervised classification techniques on web page click-through data [10] or by a text-based categorization of website's content and link structure [13]. Other research has found correlations between the demographic attributes of an author and her writing style, regarding both offline [15] and online texts [23]. Linking demographic information with *social video content*, however, has not been tackled before to the best of our knowledge.

2.4 Content-based targeting for images and videos

First attempts have also been made to employ image and video analysis for advertising. Particularly, concept detection has been used for a content-based ad targeting: in the image domain, several systems auto-detect concepts in images and personal photographs, combine them with other surrounding textual descriptions—such as user tags or text on a web page—and select a set of candidate advertisements based on this information [19,34]. For video data, this has been complemented with an analysis of the audio stream [20]. Another approach when displaying ads alongside images or videos is to localize low-saliency regions in space and time for ad placement [19,20,25]. Overall, however, while these contributions bear first promising results—indicating more accurate and less intrusive ad targeting—content-based advertising remains far from being solved. Our work follows a similar direction in a sense that we apply concept detection as well. However, our approach targets an estimation of demographic interest rather than a direct matching with ads and thus aligns more closely with common marketing practice.

3 Correlation of concepts with demographic target groups

Our goal is to estimate the distribution of user interest in a web video across different ages and genders. We split users into eight age ranges (13-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, and 75+, following the YouTube convention). Over these eight age ranges and two genders, we estimate a 16-dimensional histogram, which we will refer to as the *demographic profile* of a video clip in the following (see Fig. 2 for an illustration).

A canonical strategy would be to estimate this profile from view statistics (i.e. each time a user watches a video, the counter of his/her age/gender group is incremented). Due to privacy concerns, however, automatic access to this information is restricted. Therefore, we refer to *user comments* as a publicly accessible fall-back solution, as illustrated in Fig. 2: for all distinct users that *commented* on a video, we extract their age and gender (which we found to be specified by about 80 % of users). From this, we estimate demographic profiles as counts of comments rather than views. This information comes in lower quantities than view statistics and may introduce a certain noise, as some users may provide incorrect ages and genders. Also, the fact that we exploit commenting instead of viewing could be argued to introduce bias, for example, because posting is more popular among young users. Yet, our impression is that commenting is a good indicator for a strong engagement of users with a video and thus serves well as a measure of viewer interest.

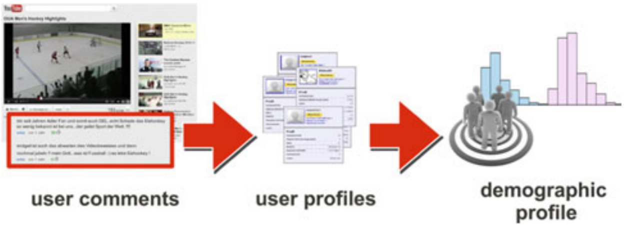

Fig. 2 To estimate the demographic viewership of a YouTube video, the age and gender across the clip's commenters are extracted and stored in a 16-dimensional histogram, which we refer to as the video's *demographic profile*

In a second step, we group the video-wise demographic profiles into distinct categories (in Sect. 4; our goal will be to automatically map videos to these categories). To do so, we apply a K-Means *clustering* on the demographic profiles and interpret the cluster centers as prototypical age and gender *distributions*, to which each video is assigned. We choose these distributions as target categories, and not the peak demographic bin (like "videos watched mostly by teenage males"). Our motivation for this is twofold: first, the overall distribution of YouTube users is strongly biased towards young and male users (for example, videos from the category "female 75+" would be extremely infrequent). Second, distributions give us a better picture of the (potentially diverse) viewership of a video than a single demographic category—think of videos whose audience covers a wide range of user ages (like "soccer" clips) as opposed to videos targeted at a strongly focused age group (like "skateboarding"): both might have male teenagers as their most frequent viewer group but still differ in wide parts of their audience, which should be reflected in our demographic categories.

In a first experiment, we downloaded a dataset of YouTube clips covering 233 semantic concepts (including objects like "car" or "cake", locations like "kitchen" or "beach", and actions like "videoblog" or "soccer"). The concept vocabulary was chosen based on earlier research in concept detection [33] and was selected with respect to observability (sufficient content on YouTube available) and feasibility of concept detection [21]. In particular, no connection to demographic viewer groups was implied when selecting these concepts.[1]

[1] For the full list of concepts, please refer to http://madm.dfki.de/demo/tubetagger.

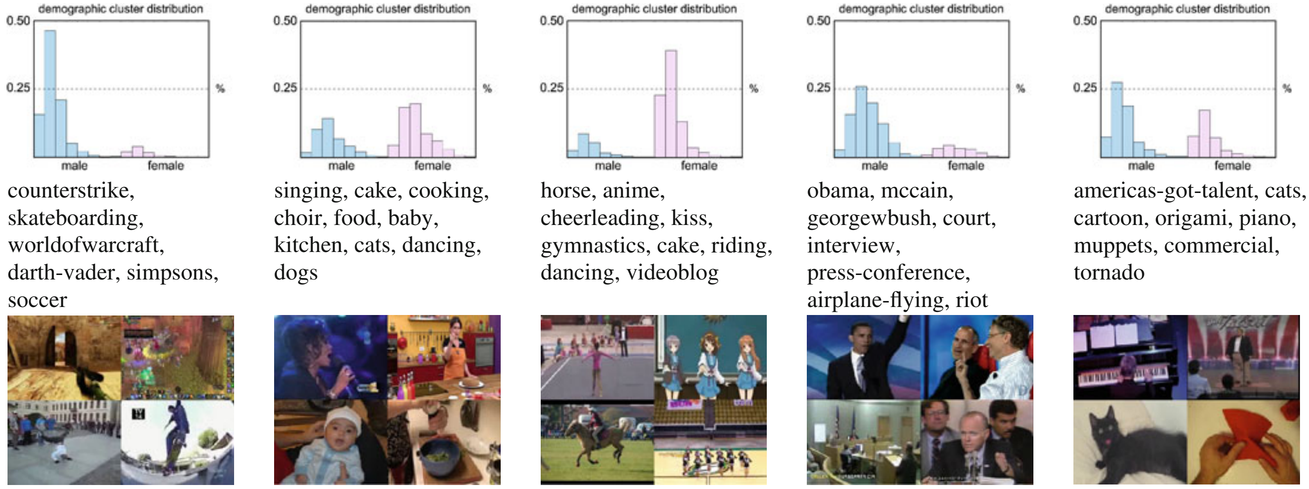

Fig. 3 We estimated five of the seven demographic clusters in a *K*-Means clustering over video commenting profiles. For each cluster, we display the cluster center as a demographic profile (*top*), the concepts with the highest number of videos in the cluster (*center*) and a few sample videos (*bottom*)

For each concept, a YouTube search for 500 videos tagged with the concept was conducted. Only videos with at least 20 user comments were kept (after this filtering, about 34 % of videos remained). We then applied *K*-Means clustering to the resulting 39,911 demographic profiles. Clustering with $K \in \{5, 7, 10\}$ was tested and based on a visual inspection $K = 7$ was chosen. After this, vector quantization was applied to all demographic profiles in the dataset—this effectively assigned each video to one of the seven demographic clusters.

Results are illustrated in Fig. 3, which displays each cluster center as a demographic histogram (top) as well as the concepts with the most videos in the cluster (center + bottom). We see that the concepts found align well with the different age distributions in the clusters: Cluster 1 (predominantly male teenage users) is dominated by computer games and youth culture, Cluster 2 (female) and Cluster 3 (teenage female) by terms like *dancing, baby, horse*, or *cheerleading*, Cluster 4 (middle-age male) by political terminology. Cluster 5 (the "kitchen sink") is closest to an "average" user distribution, covering a more diverse audience and a broader range of topics.

To quantify the correlation of concepts with certain demographic clusters, we compute the *entropy* of the distribution of a concept's videos across the demographic categories. This entropy reflects how "peaked" the distribution of a concept is, i.e. whether a concepts attracts a very specific audience profile. Figure 4 plots these entropies for all concepts (sorted ascendingly). The entropies range from 0.35 to 1.58, with a median of 1.07. An even distribution would correspond to 1.95—obviously, videos tagged with certain concepts generally tend to accumulate in certain demographic clusters. The concepts "most peaked" within certain demographic groups are "counterstrike" (0.35), "skateboarding" (0.43), "cheerleading" (0.56), "horse" (0.59), and "Mc Cain" (0.62). In contrast, the concepts attracting the broadest audience are "singing" (1.58), "kitchen" (1.56), "beach" (1.55), "choir" (1.54), and "birds" (1.51). Overall, this result indicates that concepts (particular those with low entropy) can be strong indicators for the related demographic groups.

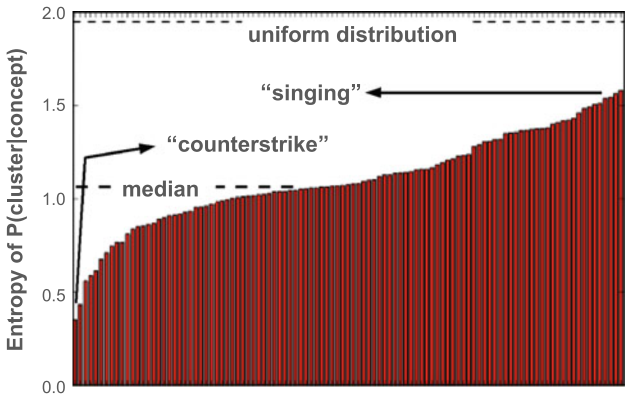

Fig. 4 For each concept, we measure the entropy of its distribution across the demographic categories. This plot shows the resulting entropy over all concepts (sorted ascendingly). Overall, these entropies are low, which indicates that concepts are strongly correlated with demographic target groups

4 Content-based demographics prediction

As was shown in Sect. 3, demographic interest is strongly connected to the topicality of video clips. Based on this fact, we present an approach called *content-based demographics prediction*, which—via a content-based *concept detection*—relates a video's content to its demographic target group.

As our focus is not on improving concept detection in general—but rather on its relation to viewer demographics—we follow a standard concept detection setup: a vocabulary of target concepts is defined, including various objects, locations, or actions (we use the same set of concepts as in Sect. 3). Concept detection is operated on keyframe level, i.e. keyframes are extracted from the video stream, represented by content-based features, and a binary classifier is first trained on a training set and then applied to held-out test videos. Thereby, we employ YouTube tags as ground truth labels: for each concept, a YouTube search is conducted, and keyframes extracted from the resulting videos serve as positive training samples. Negative samples are drawn from other concepts' videos. The resulting detectors can then be applied to previously unseen test videos and estimate probabilistic recognition scores, i.e. tags are auto-suggested. In the following, a formalization is provided on how this concept detection setup is combined with demographics prediction. For further details on the specific features and classifiers used in concept detection, please refer to Sect. 6.

Formally, our goal is to assign a video (represented by content-based features x) to one of the seven demographic clusters, $d_1, \ldots, d_7$ from Sect. 3. This demographic cluster is modeled as a random variable D, i.e. we estimate $P(D|x)$. A vocabulary of n concepts is assumed to be given. These concepts induce binary random variables $C_1, \ldots, C_n$ indicating concept presence ($C_i = 1$) or concept absence ($C_i = 0$). For each of the concepts, a concept detector has been trained on a dataset of user-tagged YouTube content to estimate a probabilistic score $P(C_i = 1|x)$ from the video representation x. This way, we obtain a vector of concept scores $P(C_n = 1|x), \ldots, P(C_n = 1|x)$.

This knowledge of concept presence is now integrated with the distribution of concepts over the different demographic clusters from Sect. 3: we use the set of all training videos in the demographic cluster j, $\mathbf{V}_j$, to compute a simple estimate for the probability that a video in cluster j shows a concept C_i:

$$P(C_i = 1|D = d_j) = \frac{1}{|\mathbf{V}_j|} \cdot |\{x \in \mathbf{V}_j | C_i(x) = 1\}| \tag{1}$$

where $C_i(x) = 1$ iff video x is tagged with concept C_i. These pieces of information—namely, semantic concepts $P(C_i = 1|x)$ and their distribution over demographic categories $P(C_i = 1|D = d_j)$—are combined by marginalizing over all possible combinations of concept appearances:

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

As evaluating this formula comes with exponential effort, we simply by assuming independence of the individual concepts and applying Bayes' rule:

$$\begin{aligned}
&\approx \sum_{c_1,c_2,\ldots,c_n\in\{0,1\}} \Bigg[\prod_{i=1}^{n} P(C_i = c_i|x)\cdot\\
&\qquad \frac{P(D = d_j)\prod_{i=1}^{n} P(C_i = c_i|D = d_j)}{\prod_{i=1}^{n} P(C_i = c_i)}\Bigg]\\
&= P(D = d_j)\cdot\prod_{i=1}^{n}\Bigg[\frac{P(C_i = 0|x)\cdot P(C_i = 0|D = d_j)}{P(C_i = 0)}\\
&\quad + \frac{P(C_i = 1|x)\cdot P(C_i = 1|D = d_j)}{P(C_i = 1)}\Bigg],
\end{aligned} \tag{2}$$

whereas simple canonical estimates are used for $P(D)$ and $P(C_i)$, based on counts of videos belonging to a certain cluster or tagged with a certain concept. fraction of training videos tagged with C_i within cluster $P(D)$. Overall, Eq. (2) provides us with a straightforward strategy to estimate the demographic profile of a clip via its concept detection results.

5 Viewer demographics as a signal for concept detection

Section 4 has introduced an approach for estimating the demographic distribution of a clip's viewership from the video's content, employing concept detection as an intermediate step. The accuracy of concept detection itself, however, is known to be far from a careful manual annotation. A common approach to boost precision is the combination of multiple heterogeneous features, including different visual aspects like local patches or color [31] as well as other modalities like the audio stream [5] or textual description accompanying the content [11]. Following this line of thought, we investigate *demographic* input signals to concept detection: instead of *estimating* demographic profiles (as in Sect. 4), we use existing ones as an additional input to disambiguate concept detection. Think of visually similar concepts (like "ice hockey" vs. "figure skating"): A content-only concept detector may remain in conflict, which can be resolved by demographic signals—for example, a video predominantly viewed by male users is more likely to show "ice hockey".

It should be noted that this approach focuses on a different application scenario than than demographics prediction (Sect. 4): While demographics *prediction* is of interest for targeting freshly uploaded videos (for which a potential future audience is to be predicted), the following approach focuses on videos for which sufficient view data exist. Still, concept detection may be of interest, for example, to improve the semantic description of the video or to localize concepts at certain scenes within.

Our approach is to feed the demographic profile of a video to concept detection, alongside conventional content-based descriptors: we assume there are two representations for a video clip, x^{demogr} and $x^{content}$. $x^{content}$ is a numerical descriptor of the video content, usually a vector of m (several hundred or thousand) dimensions. x^{demogr} is the demographic profile of the video, i.e. its 16-dimensional age/gender histogram reinterpreted as a feature vector. We propose several techniques combining these two information sources, aligning with common fusion approaches in multimedia analysis [1]:

- **Early fusion—concatenation**: The features x^{demogr} and $x^{content}$ are combined to a joint feature vector x, which is then used for classifier training and classification. As a simple combination strategy, we choose the vector concatenation $x := x^{demogr}||x^{content}$.
- **Early fusion—outer product**: as an alternative, both modalities are combined by their outer product:

$$x = x^{demogr} \otimes x^{content}$$
$$= \begin{bmatrix} x_1^{demogr} & x_1^{content} & \dots & x_1^{demogr} & x_m^{content} \\ x_2^{demogr} & x_1^{content} & \dots & x_2^{demogr} & x_m^{content} \\ & \vdots & \vdots & \ddots & \vdots \\ x_{16}^{demogr} & x_1^{content} & \dots & x_{16}^{demogr} & x_m^{content} \end{bmatrix}$$

  reinterpreted as a 16 × m-dimensional vector.
- **Early fusion with dimensionality reduction**: as content-based descriptors typically outnumber the demographic one in length, a dimensionality reduction can be applied prior to combination to balance the influence of both features. Then, $x^{content}$ is replaced with a lower-dimensional version in the above formulas.
- **Late fusion**: instead of combining feature vectors, an alternative is to train separate classifiers—one based on $x^{content}$, one on x^{demogr}—and combine their output scores, for example by a simple averaging:

$$P(C_i = 1|x) = \frac{1}{2}[P(C_i = 1|x^{content}) \\ +P(C_i = 1|x^{demogr})].$$

  As an alternative to the average, we also tested a maximum fusion.

6 Experiments

In this section, we describe quantitative results on linking demographics with concept detection on a dataset of commented YouTube videos. Section 6.5 will outline the estimation of demographic video profiles via concept detection (the approach was described in Sect. 4), Sect. 6.6 the use of demographic profiles as an additional signal for concept detection (as outlined in Sect. 5).

6.1 Dataset

The basis of the following experiments is a dataset of 35,000 YouTube clips (2,800 h of content) downloaded in 2009. Starting from the same 233-concept vocabulary as used in Sect. 3, we downloaded 150 videos per concept. All videos came from different uploaders to guarantee a high diversity of the sampled content and to avoid bias due to series of content from a single user. To improve the alignment of the downloaded content with the targeted concepts, textual queries were manually improved (like excluding the term "table" for the concept "tennis") and downloads were optionally restricted to a certain YouTube category (like "sports"). To train and test concept detection, videos were labeled according to the download (i.e. videos resulting from "tennis" downloads are labeled with the concept "tennis"). Additionally, YouTube comments from 2.2 mio. users (of which 80 % specified their age and gender) were collected for all videos. As we require reliable demographic profiles for our quantitative evaluation, we removed all videos with fewer than 20 comments and dropped concepts with less than 60 videos remaining. This reduced the vocabulary to 105 concepts and the number of videos to 14,000 (commented by about 1 mio. users) corresponding to 1,300 h of content.

To choose the videos with the most reliable demographic profiles as test videos, the clips for each concept were ranked by the number of unique users that commented on them, and the top 50 videos were chosen as test videos (5,250 overall), and the others for training concept detection (we validated in previous tests that this split by the number of comments only had a minor influence on concept detection accuracy).

6.2 Concept detection

For each clip, key-frames were extracted by a simple change detection, and concept detection was conducted on key-frame level. For each concept, a detector was trained on a held-out set of training clips. 5,000 positive and 25,000 negative key-frames were sampled per concept, whereas a keyframe counted as a positive sample for a concept c exactly if the video it was extracted from was retrieved by a YouTube search for concept c. From each test clip at most 20 random key-frames were selected, and each was fed to the 105 concept detectors. For each concept, the resulting 20 scores were combined to a joint video-level score by a simple averaging. We tested three different content-based features:

- **COLOR**: A common approach is to represent images and video frames by the distribution of their color. We choose color histograms and correllograms as a standard approach: The image is transferred to HSV color space, which is partitioned using an anisotropic binning into 10 × 6 × 5 units. Over the resulting quantized image, a color histogram and an auto-correllogram are computed (each of dimension 300), obtaining a 600-dimensional feature.

- **VISW-2000**: We also test *visual words* features, probably the most popular content-based image description technique over the past years [27]. Following standard practice, we extract local features by a dense regular sampling at multiple scales. Each region of interest is described using the SIFT feature extraction [18], obtaining 3,600 local feature vectors. These are vector-quantized to 2,000 clusters of a codebook trained previously using a K-Means clustering on half a million interest points.
- **VISW-80**: This setup is targeted at a combination of visual words features with demographic profiles. Here, a simple concatenation is bound to fail, as the much higher-dimensional visual words (2,000 dimensions) will outrule the demographic features (16 dimensions). To balance the influence of both features, add an additional dimensionality reduction step using Probabilistic Latent Semantic Analysis (PLSA) [7]. This approach replaces 2,000-dimensional histograms over visual words with 80-dimensional histograms over "topics", clusters of visual words found using a likelihood maximization with the EM algorithm.

Each of these features was fed to a Support Vector Machine (SVM) classifier [24] using a χ^2 kernel for visual words and an RBF kernel for the color features. SVM parameters were estimated using a cross-validated grid search, and the resulting scores were mapped to probabilities using the method by Lin et al. [17].

6.3 Evaluation measures

We evaluate demographics prediction (Sect. 4) as well as demographic signals (Sect. 5) quantitatively on a held-out set of test videos. To evaluate the accuracy of concept detection, we adhere to *mean average precision* as a commonly used standard measure [29] that assesses the quality of ranked retrieval lists in concept-based video search. To assess the accuracy of demographics prediction, we again choose a retrieval-based evaluation setup: for each of the seven demographic clusters d_j, all test videos x are ranked by their corresponding score $P(D = d_j|x)$. By aligning these scores with the demographic cluster assignments from Sect. 3, we compute the average precision over the video-to-cluster ranking. The average precision is again averaged over all clusters, obtaining a *Mean Average Precision* (MAP). Our rationale for choosing this retrieval-based evaluation over a classification-based one is that targeting in practice is a retrieval process:[2] Advertisers *search* for appropriate videos to place their ads with, and their satisfaction is determined by how accurately a targeting tool supports them with finding the intended target audience. Our evaluation emulates this search for demographic audience groups, and measures the accuracy of this search, effectively assessing our demographics prediction in a hypothetical demographics-based targeting tool.

6.4 Experiment 1: content-based demographics prediction

In the following, we evaluate the approach outlined in Sect. 4 that realizes an automatic mapping of videos to demographic clusters. For each of the 5,250 test videos, we estimate its demographic profile via user comments and map the video to one of the seven demographic K-Means clusters from Sect. 3. Our goal is to automatically assign the test video to its correct cluster and we use the *mean average precision* measure (see above) to assess the accuracy of this video-to-cluster assignment. As visual features, VISW-2000 (i.e. 2,000-dimensional visual word histograms) were used, which were found to give the best accuracy (more details will be provided later). Several systems were tested:

- **Random**: as a first reference, we use a random assignment of videos to demographic clusters.
- **Baseline**: as a second baseline we use a *direct* visual classification into demographic clusters, i.e. the training set is divided according to the seven clusters and for each cluster a separate 2-class SVM classifier is trained on visual features from the cluster. Applying this classifier yields scores $P(D = d_1|x), \ldots, P(D = d_7|x)$ for each test video x.
- **Marginalization**: the approach is presented in Sect. 4, which employs marginalization to integrate concept scores with the distribution of concepts over demographic clusters.
- **Hierarchical classification**: here, the marginalization outlined in Eq. (2) is replaced with an SVM classification, resulting in a two-stage process: on the first level, concept detection is applied, obtaining concept scores $P(C_1 = 1|x), \ldots, P(C_n = 1|x)$. These scores are reinterpreted as a feature vector, which is fed to a second set of seven χ^2 kernel SVMs estimating the target scores $P(D = d_1|x), \ldots, P(D = d_7|x)$. A fivefold cross-validation on the test set was used for evaluation (including the training of the second layer of SVMs).
- **Oracle**: as concept detection is usually far from accurate, in a control experiment we also test a system that replaces the concept detection scores $P(C_1{=}1|x), \ldots, P(C_n{=}1|x)$

[2] See http://www.youtube.com/t/advertising_video_targeting for YouTube's targeting tool.

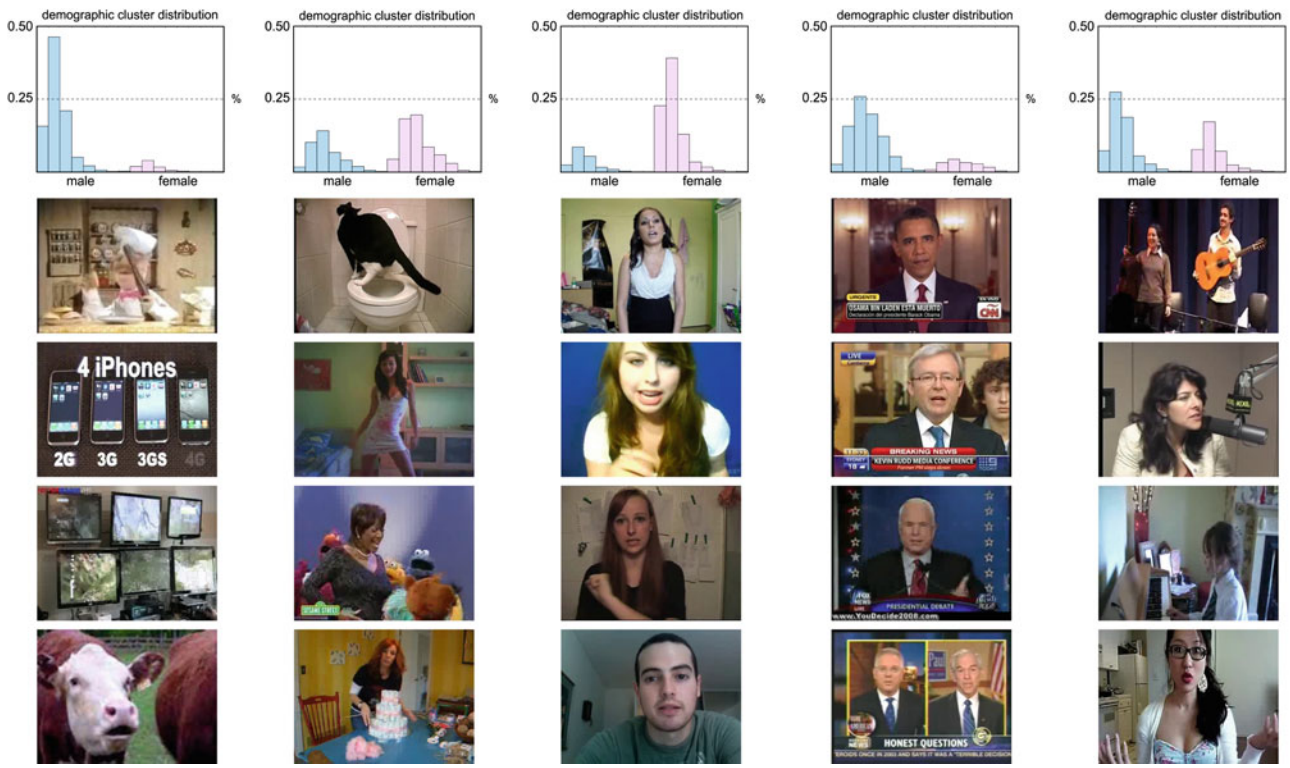

Fig. 5 Each column shows the top-ranked test videos for the corresponding demographic cluster above (using the *Marginalization* approach). These results—based only on the video content, i.e. no tags and titles were used—appear to be noisy. However, overall reasonable hits are found, like the "middle-aged male" cluster (column 4) showing mostly politics

with a binary vector indicating the true concepts according to the video's tags, i.e. $P(C_i|x) = 1$ exactly if the video x was retrieved when searching for concept C_i on YouTube. This corresponds to a hypothetical perfect concept detection. The vector of concept scores is then fed to marginalization (Eq. 2).

Figure 6 illustrates the mean average precision (MAP) for the different approaches. We see that the direct classification into clusters ("baseline", MAP 17.1 %) achieves only moderate improvements over a random assignment (MAP 14.3 %), which can be attributed to the enormous diversity of the demographic clusters: for example, the "teenage male" cluster (Fig. 3, left) contains computer games as well as outdoor skateboarding, comics, etc. Correspondingly, the results suggest that instead of modeling those highly complex demographic clusters directly, a system should rather detect semantic concepts (which are more coherent and thus feasible to discriminate) and then perform reasoning on the level of these concepts. This is confirmed by our results, as both the hierarchical classification and the marginalization approach give better accuracies, with the marginalization approach performing best (MAP 25.3 %).

While all these approaches were based on a (highly inaccurate) content-based auto-annotation of the test videos, the "oracle" run gives an indication that a much more accurate assignment would be possible if we were able to improve concept detection significantly: here, a mean average precision of 41.2 % is achieved.

Figure 5 illustrates the top-ranked videos for five of the seven demographic clusters. We see that results are noisy (see the "cow" video in the "teenage male" cluster), but that many videos seem well aligned with the interests of users in the respective clusters: the "teenage male" cluster (Column 1) also shows two videos with technical gadgets, the "young female adult" cluster (Column 2) a cat, and cake baking instructions, the "teenage female" cluster (Column 3) videoblogs, the "middle-aged male" political interviews (Column 4), and the "neutral" cluster (Column 5) music-related videos.

Confusion of demographic groups The above results have indicated that confusion between the different demographic categories is generally high. Therefore, we also study this confusion in more detail, i.e. we investigate which clusters are confused most often. We measure the confusion $\mathbf{C}_{D_i \to D_j}$ between two demographic groups D_i, D_j by averaging classification scores:

$$\frac{1}{|\mathbf{V}_i|} \cdot \sum_{x \in \mathbf{V}_i} P(D_j|x),$$

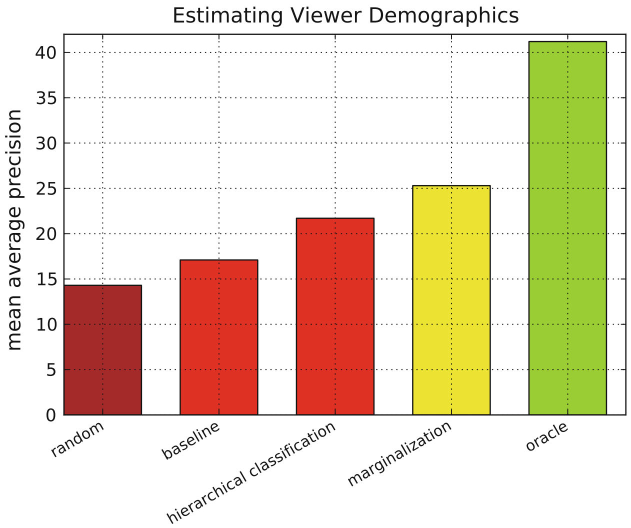

Fig. 6 Quantitative results when automatically assigning videos to demographic clusters. The marginalization approach (Sect. 4) outperforms a direct classification of demographic clusters ("baseline") and a concept-based SVM classification ("hierarchical classification"). Significant performance loss occurs due to inaccuracies of concept detection, as a control run with perfect concept detection ("oracle") indicates

i.e. $\mathbf{C}_{D_i \rightarrow D_j}$ is high if videos from cluster D_i tend to have a high classification score for D_j. Figure 7 illustrates this confusion in form of arrows pointing from each cluster C_i to the cluster $C_j (j \neq i)$ with highest confusion $\mathbf{C}_{D_i \rightarrow D_j}$. We see that using our content-based approach, each cluster is confused most often for the cluster with the *most similar* demographic distribution. This can be explained by a certain smoothness of $P(C|D)$ between demographic groups: when moving from a demographic cluster to a similar one, certain topicality shifts will occur, but interest in many topics will be similar. Also, this result is good news from a practical perspective: even if our classifier maps videos to the wrong demographic cluster, it still tends to pick a cluster with a *similar* viewership profile, i.e. targeting can still reach the right audience.

6.5 Experiment 2: comparison of modalities for demographics prediction

In a second experiment, we benchmark the content-based demographics prediction from Sect. 4 (using the marginalization approach, denoted as ***Visual*** in the following) against two other features available for social videos, namely uploader information and textual meta-data:

- **Uploader information (*uploader*)**: the first approach is based on the assumption that the uploader and the audience of videos are demographically correlated. For example, skateboarding videos—which are typically viewed by teenage male users—tend to be uploaded by users from the same demographic group. Therefore, we represent each video by a 27-dimensional indicator vector over three genders (male, female, "unknown") and nine age categories (the 8 YouTube age categories + "unknown"). This feature describes the demographic group of the video uploader. It is fed to an SVM with an RBF kernel, which was found to give higher accuracy than a χ^2 one.
- **Textual meta-data (*tags*)**: the second approach employs the textual meta-data accompanying each YouTube clip as a feature: we store all tags and all words from the clip's title, which (after lower-casing and ignoring 220 stop words as well as numbers) form a sparse indicator vector. This is fed to an SVM (again, an RBF kernel was found superior over a χ^2 one).
- **Weighted-sum fusion (*fusion*)**: finally, we also test the combination of all 3 systems (*visual*, *uploader*, *tags*) using a weighted sum fusion.

The resulting demographics predictors were trained and tested using the same data and setup as in Sect. 6.5. SVM parameters were optimized using a cross-validated grid search. Results are illustrated in Fig. 8: The visual system outperforms the uploader information, but is by itself outperformed by the text-based system. This aligns with earlier result (Fig. 6) in which a "perfect" concept detection leads to strong improvements in demographics assignment: Obviously, tags and titles—though noisy—give a better impression of a videos high-level topicality compared with a content-based concept detection.

Finally, the fusion of these three heterogeneous information sources gives a further performance gain, reaching the best mean average precision of 32.8 %. This indicates that content-based signals—though not the best individual approach—are definitely a useful input for demographics prediction. It should be kept in mind that our benchmark datasets were acquired by performing a text search for a selected set of concept names, such that (a) the selection of videos in our limited experimental setup cannot represent the full diversity of YouTube content, (b) tags are biased towards the query terms used for downloading, and (c) videos in our evaluation are relatively strongly tagged. Therefore, in a practical setting the content-based system may be less accurate (as it needs to cope with an even higher diversity of content), and so may the tags-based system (as tags in practice are less frequent and more diverse). Therefore, uploader information (though outperformed in our experiments) may still be a valuable information source in practice.

6.6 Experiment 3: demographic signal for concept detection

To evaluate the accuracy of concept detection when including demographic information as an additional signal, we apply

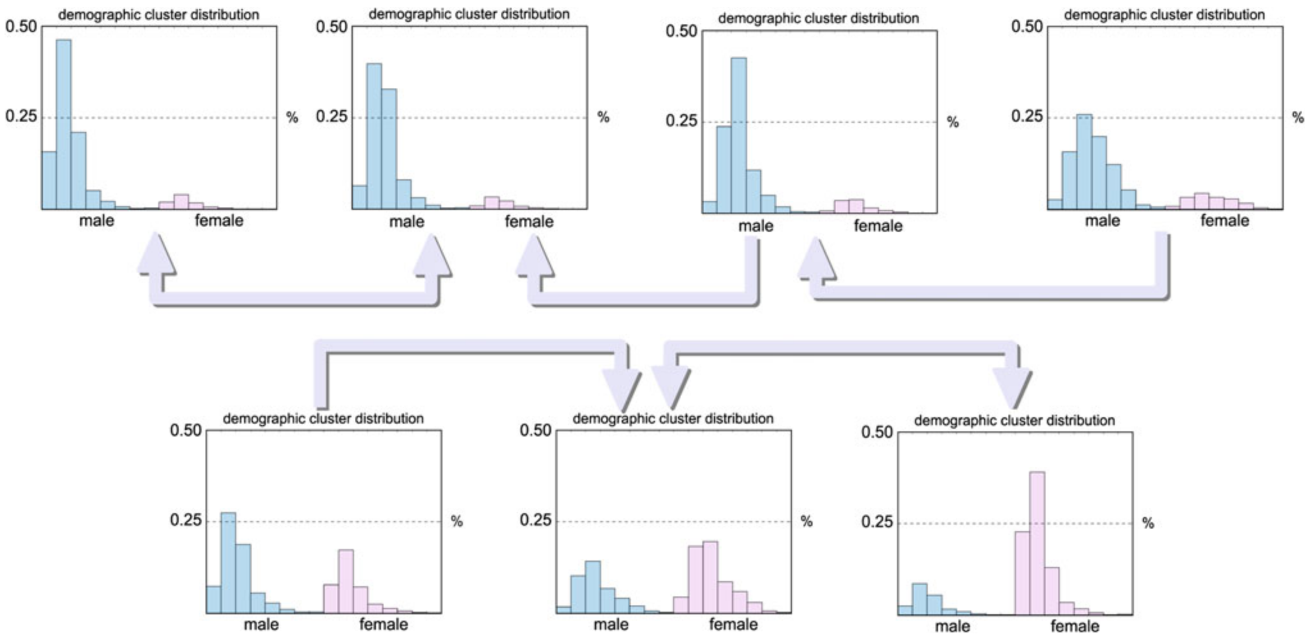

Fig. 7 Confusion between the different age categories by the visual classifier (using the marginalization approach). *An arrow* pointing from cluster C_i to C_j indicates that videos from cluster C_i are most often mapped to C_j. We see that confusion of the visual classifier correlates with demographic overlap: each cluster is confused with the (demographically) *most similar* cluster

all concept detectors on the test set and rank all 5,250 test videos for each concept. An example is illustrated in Fig. 9, where the top-ranked key-frames for the concept "surfing" are illustrated for the content-only visual words detector (left) and when including the demographic profile as an additional feature (right). We see that the content-only system gives many false positives that are visually similar to surfing (like beach scenes and panoramic landscape shots). However, by adding the demographic profile, videos less popular among young male adults are inhibited, and the overall result improves.

Quantitative results are also given in Fig. 9. Among the systems employing only a single feature (red/orange), 2,000-dimensional visual words perform best, with an average precision (AP) of 8.8 %. The demographic profile alone gives an AP of 6.5 %. When comparing both systems, the concepts for which the demographic profile was found to give the best performance were "cake" "counterstrike", "riding", "horse", and "baby" (all of them show a clear demographic profile and were rather difficult to discriminate by their content).

When combining demographic information and content in an early fusion (yellow) or a late fusion (green), we observe significant improvements. The best system—a late fusion by a simple averaging of visual score and demographic score—gives a mean average precision of 12.9 %, which corresponds to a relative improvement of 47 % over the visual-only baseline. This supports our hypothesis that demographic information can help to improve concept detection.

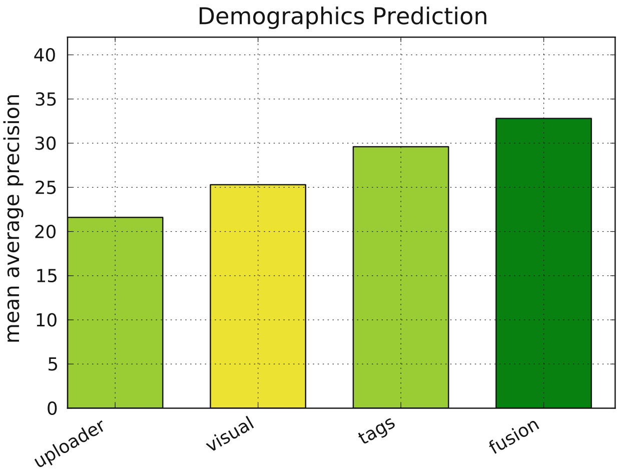

Fig. 8 Accuracy of demographics prediction when using different modalities. The highest accuracy is achieved by a combination of the content-based prediction with uploader demographics and videos' textual meta-data

7 Conclusions

In this paper we have presented a novel approach for automatic web video understanding by linking content-based concept detection with the demographic target group of video clips hosted at online platform like YouTube. Through our study both directions of this link have been investigated: first, the prediction of viewer demographics by concept detection, which provides an interesting signal for targeted adver-

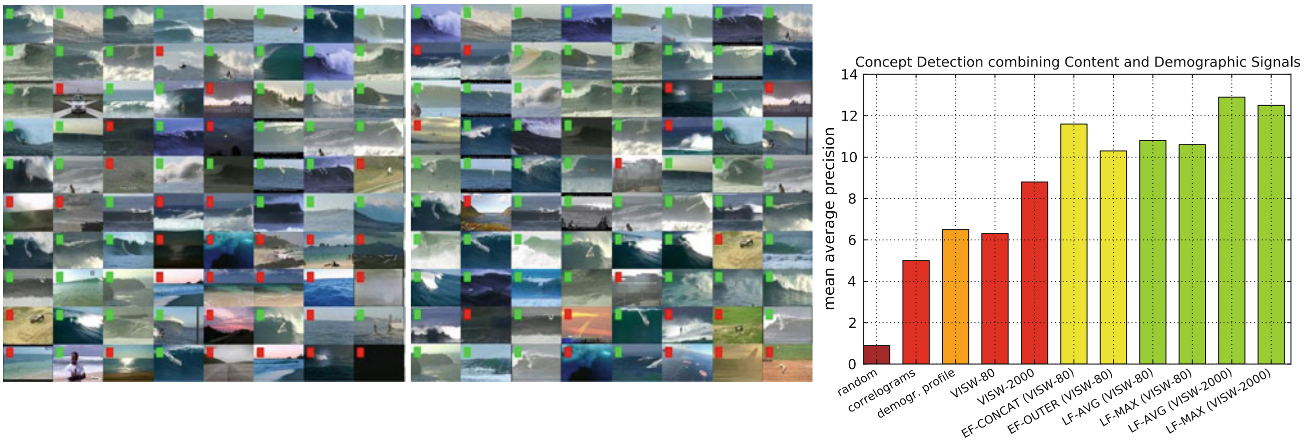

Fig. 9 Top-ranked video scenes for a “surfing” concept detector, when using only content analysis (*left*) and when combining content and demographic profiles (*center*). *The green* and *red marks* indicate concept presence and concept absence. *Right* Quantitative results of concept detection.

tising, particularly for videos found in the “long tail” of weakly annotated clips with very focused viewerships. And second, if demographic information is available, it can be utilized to improve concept detection performance significantly and therefore generate valuable meta-data for web video.

In experiments on 14,000 YouTube videos commented by 1 mio. users we showed that a robust demographic group prediction for videos is possible but limited by the accuracy of the concept detection system. To overcome this, fusion with further modalities like tag information and demographics of the video uploader has been introduced and demonstrated to improve overall system performance significantly. Additionally, we demonstrated that the usage of demographic information of a video clip itself as an additional signal for concept detection leads to a relative performance improvement of 47 %.

Outlying further research in this direction, an additional in-depth profiling of a user might be of value to predict the demographic group of a video clip he uploaded. This might include users’ viewing preferences, ratings, or subscribed channels. Thinking about demographic information as a helpful additional signal for concept detection, a step closer to the user would mean to personalize concept detection to the users’ upload history, as users tend to upload videos similar to the content they uploaded in the past.

Acknowledgments This work was sponsored by the Google Research Awards Program. We thank Luciano Sbaiz from Google for his valuable input on the topic.

References

1. Atrey P, Hossain M, El Saddik A, Kankanhalli M (2010) Multimodal fusion for multimedia analysis: a survey. Multimed Syst 16(6):345–379
2. Bozios T, Lekakos G, Skoularidou V (2001) Advanced techniques for personalized advertising in a digital TV environment: the IMedia system. In: Proceedings of the eBusiness and eWork conference
3. Chen Y, Pavlov D, Canny JF (2010) Behavioral targeting: The art of scaling up simple algorithms. ACM Trans. Knowl. Discov. Data 4(4):17:1–17:31. doi:10.1145/1857947.1857949
4. ComScore August 2011 U.S. Online Video Rankings. http://www.nytimes.com/2010/04/07/business/media/07adco.html (retrieved: Oct’12), April 2010
5. Cao L et al (2011) IBM Research and Columbia University TRECVID-2011 multimedia event detection (MED) system. In: Proceedings of the TRECVID workshop. http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/ibm.pdf
6. Everingham M, Van Gool L, Williams C, Winn J, Zisserman A (2010) The Pascal Visual Object Classes (VOC) Challenge. Int. Journal of Computer Vision 88(2):303–338
7. Hofmann T (2001) Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42:177–196
8. Hollis (2005) Ten years of learning on how online advertising builds brands. Advert Res 45:255–268
9. Hrishikesh A, Toderici G, Yagnik J (2009) Video2Text: learning to annotate video content. In: Proceedings of the workshop on internet multimedia, mining
10. Hu J, Zeng H-J, Li H, Niu C, Chen Z (2007) Demographic prediction based on user’s browsing behavior. In: Proceeidngs of WWW, pp 151–160
11. Huurnink B, Snoek C, de Rijke M, Smeulders A (2012) Content-Based Analysis Improves Audiovisual Archive Retrieval. Multimedia, IEEE Transactions on 14(4):1166–1178
12. Jansen BJ, Mullen T (2008) Sponsored Search: An Overview of the Concept, History, and Technology. IJEB 6(2):114–131
13. Kabbur S, Han E-H, Karypis G (2010) Content-based methods for predicting web-site demographic attributes. In: Proceedings of ICDM, pp 863–868
14. Kennedy L, Chang S-F, Kozintsev I (2006) To search or to label?: predicting the performance of search-based automatic image classifiers. In: Workshop multimedia, information retrieval
15. Koppel M, Argamon S, Shimoni AR (2002) Automatically categorizing written texts by author gender. Lit Linguist Comput 17(4):401–412
16. Li X, Snoek C, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. In: Proceedings of MIR, pp 180–187

17. Lin H-T, Lin C-J, Weng R (2007) A Note on Platt's Probabilistic Outputs for Support Vector Machines. Mach. Learn. 68(3): 267–276
18. Lowe D (2004) Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 60(2):91–110
19. Mei T, Hua X-S, Li S (2008) Contextual in-image advertising. In: Proceedings of ACM multimedia, pp 439–448
20. Mei T, Hua X-S, Li S (2009) VideoSense: A Contextual In-video Advertising System. IEEE Trans. Cir. and Sys. for Video Technol 19:1866–1879
21. Naphade M, Smith J, Tesic J, Chang S, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE MultiMed 13(3):86–91
22. Over P, Awad G, Fiscus J, Antonishek B, Smeaton AF, Kraaij W, Quenot G (2010) TRECVID 2010-an overview of the goals, tasks, data. Evaluation mechanisms and metrics. In: Proceedings of TRECVID workshop
23. Schler J, Koppel M, Argamon S, Pennebaker J (2006) Effects of age and gender on blogging. In: Proceedings of AAAI spring symposium on computational approaches for analyzing weblogs
24. Schölkopf B, Smola A (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge
25. Sengamedu SH, Sawant N, Wadhwa S (2007) vADeo: video advertising system. In: Proceedings of ACM multimedia, pp 455–456
26. Silversmith D (2011) Google losing up to 1.65M a day on YouTube. internetevolution.com (retrieved: December 2011)
27. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proceedings of the international conference computer vision, pp 1470–1477
28. Smeaton A (2005) Large scale evaluations of multimedia information retrieval: the TRECVid experience. In: Proceedings of CIVR, pp 11–17
29. Smeaton A, Over P, Kraaij W (2009) High-level feature detection from video in TRECVid: a 5-year retrospective of achievements. Multimed Content Anal, pp 1–24
30. Snoek C, Worring M (2009) Concept-based video retrieval. Found Trends Inf Retr 4(2):215–322. http://www.science.uva.nl/research/publications/2009/SnoekFTIR2009
31. Snoek C et al (2011) The mediaMill TRECVID 2011 semantic video search engine. In: Proceedings of TRECVID workshop (unreviewed workshop paper)
32. Torresani L, Szummer M, Fitzgibbon A (2010) Efficient object category recognition using classemes. Comput Vis ECCV 2010: 776–789
33. Ulges A, Koch M, Borth D, Breuel T (2009) TubeTagger-YouTube-based concept detection. In: Proceedings of workshop on internet multimedia, mining
34. Wang X-J, Yu M, Zhang L, Cai R, Argo W-YMa (2009) Intelligent advertising by mining a user's interest from his photo collections. In: Proceedings of KDD workshop on data mining and audience intelligence for advertising, pp 18–26
35. Wesch M (2008) An anthropological introduction to YouTube. http://www.youtube.com/watch?v=TPAO-lZ4_hU (retrieved: March 2010)
36. Wu X, Bolivar A (2008) Keyword extraction for contextual advertisement. In: Proceedings of the 17th international conference on, World Wide Web, pp 1195–1196
37. Yan J, Liu N, Wang G, Zhang W, Jiang Y, Chen Z (2009) How much can behavioral targeting help online advertising? In: Proceedings of the 18th international conference on, World wide web, pp 261–270
38. Yan J et al (2009) How much can behavioral targeting help online advertising? In: Proceedings of WWW, pp 261–270
39. Yang J, Hauptmann A (2008) (Un)Reliability of video concept detection. In: Proceedings of the international conference image and video retrieval, pp 85–94
40. Yih W-t, Goodman J, Carvalho VR (2006) Finding advertising keywords on web pages. In: Proceedings of WWW, pp 213–222
41. YouTube Press Statistics. http://youtube.com/t/press_statistics (retrieved: Mar'12)
42. Zelnik-Manor L, Zanetti S, Perona P (2008) A walk through the web's video clips. In: Proceedings of first internet vision workshop