Linking Visual Concept Detection with Viewer Demographics

Adrian Ulges
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
adrian.ulges@dfki.de

Markus Koch
University of Kaiserslautern
Kaiserslautern, Germany
m_koch@cs.uni-kl.de

Damian Borth
University of Kaiserslautern and DFKI
Kaiserslautern, Germany
damian.borth@dfki.de

ABSTRACT

The estimation of demographic target groups for web videos – with applications in ad targeting – poses a challenging problem, as the textual description and view statistics available for many clips is extremely sparse. Therefore, the goal of this paper is to link a clip's popularity across different viewer *ages and genders* on the one hand with the video *content* on the other: Employing user comments and user profiles on YouTube, we show that there is a strong correlation between demographic target groups and semantic concepts appearing in the video (like "teenage male" and "skateboarding"). Based on this observation, we suggest two approaches: First, the demographic target group of a clip is predicted automatically via a content-based concept detection. Second, should sufficient view statistics already give a good impression of a video's audience, we show that this information can serve as a valuable additional signal to disambiguate concept detection.

Our experimental results on a dataset of 14,000 YouTube clips commented by 1 mio. users show that – though content-based viewership estimation is a challenging problem – suitable demographic groups can be suggested by concept detection. Also, a combination with demographic information as an additional signal leads to relative improvements of concept detection accuracy by 47%.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Indexing

General Terms

Algorithms, Measurement, Experimentation

Keywords

video analysis, concept detection, demographics estimation

1. INTRODUCTION

Over the last years, web video portals (like YouTube, Blinkx, Vimeo, etc.) have experienced a tremendous growth in both the volume of content shared and the size of their user communities. YouTube alone, the market leader in this area, sees 60 hours of video uploaded every minute, serves 4 billion views a day, and is visited by 800 million unique users each month [29]. Key issues to web video services are an effective video search and recommendation to guide visitors to the content they are interested in. Also, advertising is a vital issue, as serving videos 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 [21]). To cover these costs, web videos are linked with a variety of ads, some text-based [10], others combined with the video content (for example, by placing an ad beside or within the current clip, or by playing a pre-roll or post-roll spot).

An important issue with web video advertising is *targeting*, i.e. the selection of appropriate ads to be displayed for a certain user / with a certain video. Targeting can follow different strategies – prominent ones are a profiling of the uploader's or viewer's personal interests (*behavioral targeting* [9]), a modeling of the semantic context (*contextual advertising* [28]), or an estimation of *viewer demographics*. Particularly, demographic profiling has been applied extensively in targeting before, based on the assumption that users of different ages and genders show a specific interest in certain products. Many advertisers define the target groups for their products and services in terms of demographic attributes like age, gender, education and income [6], and targeting systems employ such demographic information [2].

A key challenge to targeting (as well as to search and recommendation) is that the information that current approaches ground on is often sparse: Imagine a video that carries no meaningful title or tags and has just been freshly uploaded – obviously, it is difficult to predict which kind of audience the video might be targeted at. Correspondingly, a large number of clips is missed by video targeting (YouTube has been estimated to monetize only 14% of its views [29]).

To overcome the sparseness of meta-data, the strategy followed in this paper will be to link viewership demographics (more precisely, age and gender) with the *content* of a video. Though the accuracy of automatic content analysis is limited, it 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 is available (if the video has just recently been uploaded), and neither is a user profile of the viewer (if not logged in) or of the uploader (if no

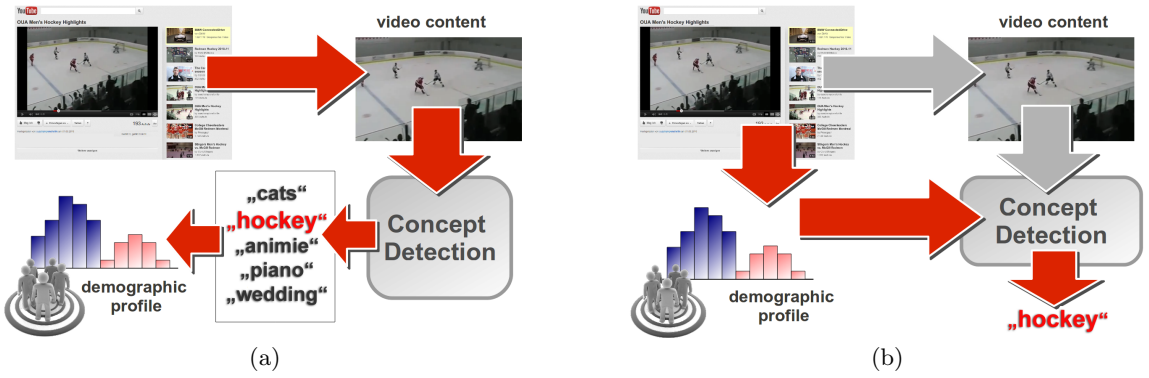

Figure 1: We study two strategies for linking viewer demographics with concept detection – *(a)* Using concept detection, we detect semantic concepts appearing in the video. The fact that these are often linked with certain demographic user groups allows us to estimate the demographic profile of a video (i.e. its popularity across different ages and genders). *(b)* 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.

age and gender were specified). We present a system that automatically predicts the semantic concepts appearing in a video, together with the popularity of a clip across different demographics. The core component forms a content-based *concept detection* engine [25] that automatically mines web clips for objects, locations and actions appearing in them. Our key contribution is that we combine this concept detection with information on video popularity across different ages and genders: 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. Based on this observation, we have implemented two strategies:

1. **Concept detection for inferring viewer demographics**: We present an approach that applies concept detection and - based on the resulting concept scores - infers the popularity of a clip across different viewer ages and genders (an illustration is given in Figure 1(a)). 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). This approach is particularly interesting for freshly uploaded videos (with limited view history), or to detect scenes within longer clips that are of particular interest to different viewer groups (imagine a video of a trip to Paris showing both museums and nightlife).
2. **Viewership demographics as a signal for concept detection**: On the other hand, if a video has already been viewed and commented on extensively, its demographic profile can serve as a signal for concept detection. As illustrated in Figure 1(b), we apply this information alongside traditional content-based descriptors, helping concept detectors to disambiguate (think of visually similar concepts that attract different audiences, like "ice hockey" vs. "figure skating").

We present experimental results on a dataset of 14,000 clips from YouTube (1,300 hours 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 (outperforming a direct visual classification of demographic categories as a baseline), and that the accuracy of concept detection can be improved significantly by joining in demographic profiles.

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

2. RELATED WORK

In the following, research related to our work will be outlined, including concept detection in general, content analysis for advertising, and the estimation of viewer demographics. For ad targeting strategies in general, which is beyond the scope of this work, please refer to [10, 9, 28].

Video Concept Detection: The challenge of automatically detecting semantic concepts such as objects, locations, and activities in video streams — referred to as *video annotation* [7], *concept detection* [27], or *high-level feature extraction* [22] — has been subject to extensive study over the last decade. In benchmarks like TRECVID [22] or the PASCAL visual object challenge [4], the research community has investigated a variety of features and statistical models – please refer to [23] for a survey.

Originally, research in the field has focused on expert-defined tag vocabularies and training data, which are limited in scalability and flexibility. More recent approaches have therefore turned towards portals like Flickr and YouTube as information sources for visual learning. Here, concept detection sys-

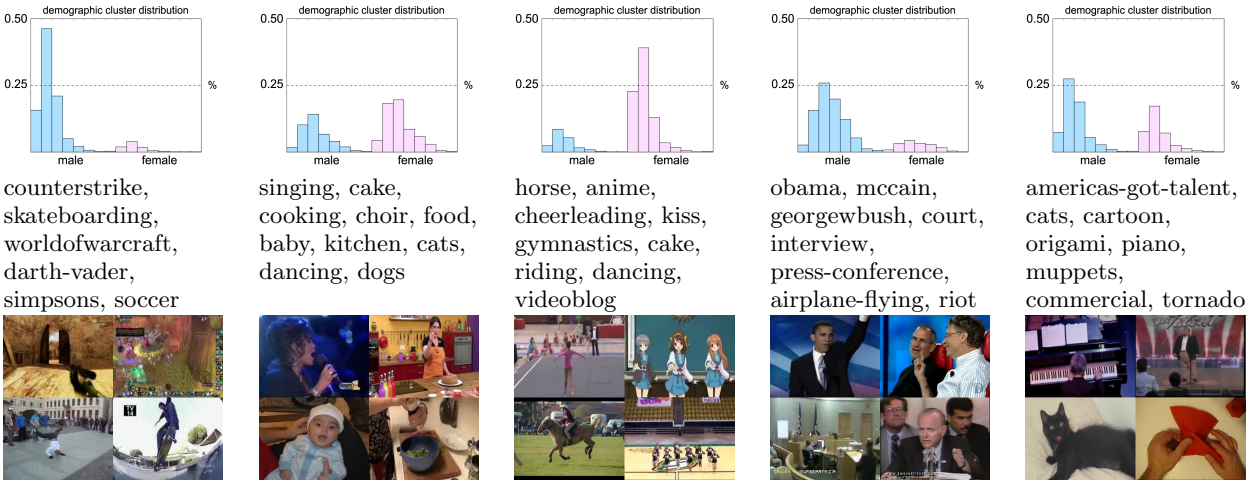

Figure 2: Five of the seven demographic clusters we estimated in a K-Means clustering over video commenting profiles. For each cluster, we display the cluster center as a demographic profile (top) and the concepts with the highest number of videos in the cluster.

tems employ user-generated tags as an alternative to expert training labels [12, 25, 30]. Key research issues include the adaptation to weak label information [13, 24] and the automatic selection of tag vocabularies [7]. 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. Our approach also bears similarities to Multimedia Event Detection [18], where high-level semantics provided by concept detection systems are applied to model further layers of abstraction [3].

Content-based Advertising: 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 [16, 26]. For video data, this has been complemented with an analysis of the audio stream [17]. Another approach when displaying ads alongside images or videos is to localize non-intrusive regions in space and time for ad placement [16, 17, 20]. Overall, however, while these contributions bear first promising results – indicating higher user satisfaction and more accurate ad targeting – content-based advertising remains far from 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.

Demographics Estimation: The automatic prediction of users' demographic attributes has been studied for conventional web browsing, mainly by training supervised classification techniques on web page click-through data [8] or by a text-based categorization of website's content and link structure [11]. Linking demographic information with *video content*, however, has not been tackled before to the best of our knowledge.

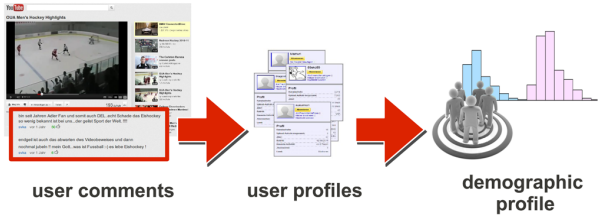

Figure 3: To estimate the demographic viewership of a YouTube video, the age and gender across a video's commenters is extracted and stored in a 16-dimensional histogram, which we refer to as the video's *demographic profile*.

3. ESTIMATING DEMOGRAPHIC PROFILES FOR VIDEO CLIPS

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 a *demographic profile* in the following (check Figure 3 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 fall-back solution, as illustrated in Figure 3: For all distinct users that *commented* on a video, we extract their age and gender (which are available for 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 also introduces a certain noise, as some users may provide incorrect ages and genders. Comments can also be expected to introduce some bias, as

they are 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 user interest.

In a second step, we group the video-wise demographic profiles into distinct categories (in Section 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. Our motivation for this is that distributions give us a better picture of the (potentially diverse) viewership of a video – 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").

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") – for more details, see [25]. 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. We then applied K-Means clustering to the resulting 39,911 demographic profiles. Cluster numbers of K $\in \{5, 7, 10\}$ were tried, and based on a visual inspection K=7 was chosen. Results are illustrated in Figure 2, which displays each cluster center as a demographic histogram (top) and 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, Clusters 2 (young female) and 3 (teenage female) by terms like *dancing*, *baby*, *horse*, or *cheerleading*, Cluster 4 (middle-aged male) by political terminology. Cluster 5 (the "kitchen sink") covers a more diverse audience and a broader range of topics.

To quantify the correlation of concepts with certain demographic clusters, we applied vector quantization to all demographic profiles in the dataset, effectively assigning each video to one of the seven clusters in Figure 2. For each concept, we used the *entropy* to measure how "peaked" the distribution of the concept's videos across the demographic clusters is. These concept entropies range from 0.35 ("counterstrike") to 1.58 ("singing"), with a median of 1.07. An even distribution would correspond to 1.95 – obviously, videos tagged with certain concepts tend to accumulate in certain demographic clusters. Particularly, if we can detect low-entropy concepts, we expect those to be strong indicators for certain demographic groups.

4. ESTIMATING VIEWER DEMOGRAPHICS BY CONCEPT DETECTION

In the following, we will use the key observation of Section 3 – namely that video commenter's age and gender are correlated with semantic concepts appearing in the video – to automatically estimate a video's demographic viewership directly from its content. Our goal is to assign a video (represented by content-based features x) to one of the seven demographic clusters, $d_1, .., d_7$. This demographic cluster is modeled as a random variable D, i.e. we estimate $P(D|x)$. To do so, we apply automatic concept detection [25]: A vocabulary of n concepts is assumed to be given. These concepts induce binary random variables $C_1, .., C_n$ 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: A binary classification problem is formulated for each concept, in which the classifier learns to "auto-tag" videos with the concept. This way, we obtain a vector of concept scores $P(C_n = 1|x)$, .., $P(C_n = 1|x)$.

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

$$P(C_i = 1|D = d_j) = \frac{1}{|\mathcal{T}_j|} \cdot |\{x \in \mathcal{T}_j | C_i(x) = 1\}| \quad (1)$$

where $C_i(x)$ denotes the presence of concept C_i in video x. These two information sources – 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:

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

Assuming independence of the individual concepts and applying Bayes' rule, we can rewrite this as:

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

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. Overall, Equation (2) provides us with a simple 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 content, by employing concept detection as an intermediate step. The accuracy of concept detection itself, however, is known to be far from a careful manual annotation. Therefore, this section introduces an inverse approach to the one in Section 4: Instead of estimating demographic profiles (for example, for freshly uploaded videos), we use existing ones as an additional input.

We exploit the fact that for many videos sufficient view data exist. Still, we may be interested in applying concept detection, for example to improve the semantic description of the video or to localize concepts at certain scenes within. What makes concept detection challenging is that target vocabularies contain many visually similar concepts (think of "ice hockey" vs. "figure skating"). Such conflicts may be resolved by demographic signals – for example, if a video is predominantly viewed by male users it is more likely to show "ice hockey". Thus, our idea 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 $x := x^{demogr}||x^{content}$ (note that – as the content-based descriptor outnumbers the demographic one in length, a dimensionality reduction may be applied to $x^{content}$ prior to combination).

- **Early fusion – outer product**: Both modalities are combined by their outer product:

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

  reinterpreted as a $16 \times m$-dimensional vector (to limit the dimensionality of x, we may again consider a prior dimensionality reduction of $x^{content}$).

- **Late fusion – combination**: 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:

$$\begin{aligned} P(C_i = 1|x) &= \frac{1}{2}\left[P(C_i = 1|x^{content}) \right. \\ &\quad \left. + P(C_i = 1|x^{demogr})\right]. \end{aligned}$$

  As an alternative to the average, we also tested the maximum and the sum of both.

6. EXPERIMENTS

In this section, we describe quantitative results on linking demographics with concept detection on a dataset of commented YouTube videos. Section 6.1 will outline the estimation of demographic video profiles via concept detection (the approach was described in Section 4), Section 6.2 the use of demographic profiles as an additional signal for concept detection (as outlined in Section 5).

Dataset: The basis of the following experiments is a dataset of 35,000 YouTube clips (2,800 hrs. of content) downloaded in 2009. Starting from the same 233-concept vocabulary as used in Section 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 hours 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), 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).

Concept Detection For each clip, key-frames are 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). From each test video, at most 20 random key-frames were selected and each is fed to the 105 concept detectors. The resulting 20 scores were combined to a joint video-level score by a simple averaging. We tested three content-based features:

- **COLOR**: a 600-dimensional color feature, consisting of an HSV histogram and HSV auto-correlogram of 300 dimensions each.

- **VISW-2000**: Following the common *visual words* approach, these features are obtained by a regular multi-scale sampling of about 3,600 SIFT features [15], vector-quantized to 2,000 clusters using K-Means.

- **VISW-80**: To balance the influence of the $2,000$-dimensional visual words features compared to the demographic histograms, we apply a dimensionality reduction to 80 dimensions using Probabilistic Latent Semantic Analysis (PLSA) [5].

These features were fed to a Support Vector Machine (SVM) classifier [19] 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. [14].

Figure 4: Each column shows the top-ranked test videos for the corresponding demographic cluster above. These results *Marginalization* approach – which employs only the video content, i.e. no tags and titles were used – appear to be noisy but several reasonable hits are found, like the "middle-aged male" cluster (column 4) showing mostly politics.

6.1 Experiment 1: Estimating Viewer Demographics

In the following, we evaluate the approach outlined in Section 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 Section 3. Our goal is to automatically assign the test video to its correct cluster — we measure the accuracy of this video-to-cluster assignment using *mean average precision*, i.e. for each cluster all test videos are ranked by their corresponding score and the average precision over this video-to-cluster ranking is computed (which is again averaged over all clusters). As visual features, VISW-2000 (i.e. 2000-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 baseline, 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 $P(D = d_1|x), .., P(D = d_7|x)$ for each test video x.
- **Marginalization**: Our approach as presented in Section 4, which employs marginalization to integrate concept scores with the distribution of concepts over demographic clusters.
- **Hierarchical classification**: Here, the marginalization outlined in Equation (2) is replaced with an SVM classification, resulting in a two-stage process: On the first level, concept detection is applied, obtaining concept scores $P(C_1 = 1|x), .., P(C_n = 1|x)$. 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 $P(D = d_1|x), .., P(D = d_7|x)$. The training of this second set of SVMs was done using a 5-fold cross-validation on the test set.
- **Oracle**: As concept detection is usually far from accurate, in a control experiment we also test a system that replaces the concept detection scores $P(C_1 = 1|x), .., P(C_n = 1|x)$ with a binary vector indicating the true concepts according to the video's tags (which would correspond to a perfect concept detection). This vector is fed to marginalization (Equation (2)).

Figure 5 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 (Figure 2, 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 is more feasible) and then perform reasoning on the level of these concepts. This is confirmed by our results, as both the hierarchical classifi-

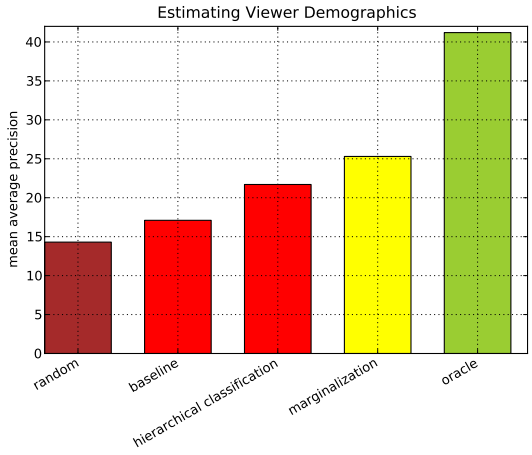

Figure 5: Quantitative results when automatically assigning videos to demographic categories. The marginalization approach (Section 4) outperforms a direct classification of demographic clusters ("baseline") and a concept-based SVM classification ("hierarchical classification"). Significant performance loss occurs due to inaccuracies of concept detection, as a control run with perfect concept detection ("oracle") indicates.

cation 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 is possible by improving concept detection, or if high-quality tags for a video are available. Here, a mean average precision of 41.2% is achieved.
Figure 4 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-age male" political interviews (Column 4), and the "neutral" cluster (Column 5) music-related videos. Overall, content-based demographic mapping (though far from accurate) may form an interesting input for ad targeting, where – given the huge number of videos and view events – signals on potential user interest are very useful.

6.2 Experiment 2: Viewer Demographics as a Signal for Concept Detection

To evaluate the accuracy of concept detection when including demographic information as an additional signal, we apply all concept detectors on the test set, rank all 5,250 test videos for each concept, and compute the mean average precision, which can be considered a standard approach. An example is illustrated in Figure 6, 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 Figure 6. Among the systems employing only a single feature (red/orange), 2000-dimensional visual words perform best (AP 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-game", "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.

7. CONCLUSIONS

We have presented an approach for automatic web video understanding that links content-based concept detection with the demographic target group of video clips. Both directions of this link have been investigated: On the one hand, the estimation of viewer demographics by concept detection makes an interesting signal for targeted advertising, particularly for the "long tail" of weakly annotated clips with very focused viewerships. On the other hand, if demographic information is available, it can improve concept detection significantly and thus help to improve web video meta-data. Following this line of research further, there is plenty of opportunity for improvement, particularly by including further signals like tag information (which might increase the accuracy of demographics estimation significantly as indicated by the "oracle" run in Figure 5) or by a more extensive user profiling[1].

8. REFERENCES

[1] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli. Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Systems*, 16(6):345–379, 2010.
[2] T. Bozios, G. Lekakos, and V. Skoularidou. Advanced Techniques for Personalized Advertising in a Digital TV Environment: The IMedia System. In *Proc. of the eBusiness and eWork Conference*, 2001.
[3] L. C. et al. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System. In *Proc. TRECVID Workshop*, 2011.
[4] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Int. Journal of Computer Vision*, 88(2):303–338, 2010.
[5] T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196, 2001.

[1]This work was sponsored by the Google Research Awards Program.

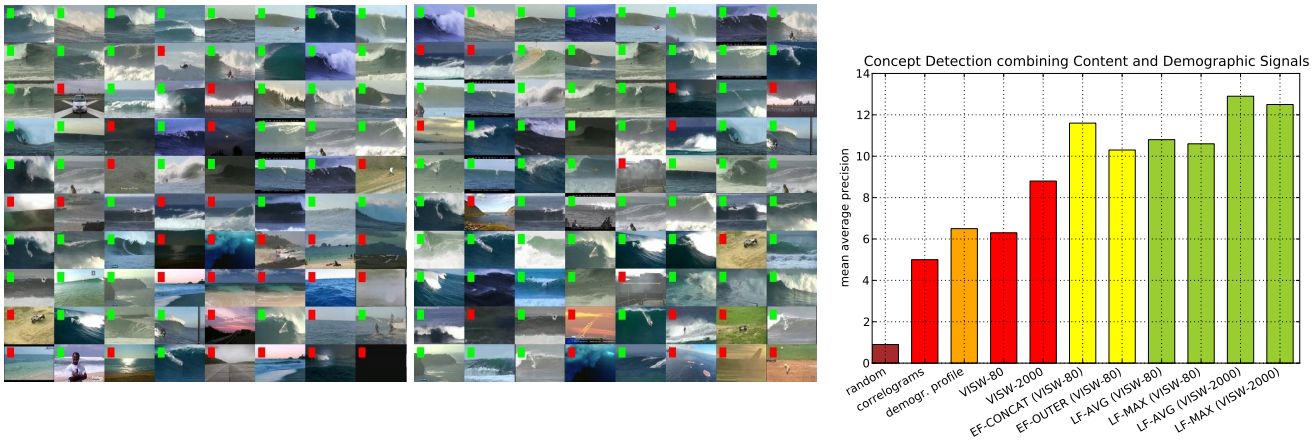

Figure 6: Top-ranked video scenes for a "surfing" concept detector, when using only content analysis (left) and when combining content and demographic profiles (center). The green and red marks indicate concept presence and concept absence. Right: Quantitative results of concept detection.

[6] Hollis. Ten Years of Learning on How Online Advertising Builds Brands. *Advertising Research*, 45:255–268, 2005.

[7] A. Hrishikesh, G. Toderici, and J. Yagnik. Video2Text: Learning to Annotate Video Content. In *Proc. Workshop on Internet Multim. Mining*, 2009.

[8] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic Prediction based on User's Browsing Behavior. In *Proc. WWW*, pages 151–160, 2007.

[9] J. Yan et al. How much can Behavioral Targeting help Online Advertising? In *Proc. WWW*, pages 261–270, 2009.

[10] B. J. Jansen and T. Mullen. Sponsored Search: An Overview of the Concept, History, and Technology. *IJEB*, 6(2):114–131, 2008.

[11] S. Kabbur, E.-H. Han, and G. Karypis. Content-Based Methods for Predicting Web-Site Demographic Attributes. In *Proc. ICDM*, pages 863–868, 2010.

[12] L. Kennedy, S.-F. Chang, and I. Kozintsev. To Search or to Label?: Predicting the Performance of Search-based Automatic Image Classifiers. In *Workshop Multimedia Information Retrieval*, 2006.

[13] X. Li, C. Snoek, and M. Worring. Learning Tag Relevance by Neighbor Voting for Social Image Retrieval. In *Proc. MIR*, pages 180–187, 2008.

[14] H.-T. Lin, C.-J. Lin, and R. Weng. A Note on Platt's Probabilistic Outputs for Support Vector Machines. *Mach. Learn.*, 68(3):267–276, 2007.

[15] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[16] T. Mei, X.-S. Hua, and S. Li. Contextual In-Image Advertising. In *Proc. ACM Multimedia*, pages 439–448, 2008.

[17] T. Mei, X.-S. Hua, and S. Li. VideoSense: A Contextual In-video Advertising System. *IEEE Trans. Cir. and Sys. for Video Technol.*, 19:1866–1879, 2009.

[18] P. Over, G. Awad, J. Fiscus, B. Antonishek, A. F. Smeaton, W. Kraaij, and G. Quenot. TRECVID 2010 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. TRECVID Workshop*, 2010.

[19] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. 2001.

[20] S. H. Sengamedu, N. Sawant, and S. Wadhwa. vADeo: Video Advertising System. In *Proc. ACM Multimedia*, pages 455–456, 2007.

[21] D. Silversmith. Google Losing up to $ 1.65M a Day on YouTube. available from internetevolution.com (retrieved: December 2011).

[22] A. Smeaton. Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience. In *Proc. CIVR*, pages 11–17, 2005.

[23] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[24] A. Ulges, D. Borth, and T. Breuel. Visual Concept Learning from Weakly Labeled Web Videos. In *Video Search and Mining*. Springer-Verlag, 2010.

[25] A. Ulges, M. Koch, D. Borth, and T. Breuel. TubeTagger – YouTube-based Concept Detection. In *Proc. Workshop on Internet Multim. Mining*, 2009.

[26] X.-J. Wang, M. Yu, L. Zhang, R. Cai, and W.-Y. Ma. Argo: Intelligent Advertising by Mining a User's Interest from his Photo Collections. In *Proc. KDD Workshop on Data Mining and Audience Intelligence for Advertising*, pages 18–26, 2009.

[27] J. Yang and A. Hauptmann. (Un)Reliability of Video Concept Detection. In *Proc. Int. Conf. Image and Video Retrieval*, pages 85–94, 2008.

[28] W.-t. Yih, J. Goodman, and V. R. Carvalho. Finding Advertising Keywords on Web Pages. In *Proc. WWW*, pages 213–222, 2006.

[29] YouTube Press Statistics. available from youtube.com/t/press_statistics (retrieved: Mar'12).

[30] L. Zelnik-Manor, S. Zanetti, and P. Perona. A Walk Through the Web's Video Clips. In *Proc. First Internet Vision Workshop*, 2008.