TubeFiler – an Automatic Web Video Categorizer*

Damian Borth†,
Jörn Hees,
Markus Koch
Univ. of Kaiserslautern
Kaiserslautern, Germany

Adrian Ulges,
Christian Schulze,
Thomas Breuel
DFKI
Kaiserslautern, Germany

Roberto Paredes
DSIC
Universidad Politécnica
Valencia, Spain

ABSTRACT

While hierarchies are powerful tools for organizing content in other application areas, current web video platforms offer only limited support for a taxonomy-based browsing. To overcome this limitation, we present a framework called *TubeFiler*. Its two key features are an automatic multimodal categorization of videos into a genre hierarchy, and a support of additional fine-grained hierarchy levels based on unsupervised learning. We present experimental results on real-world YouTube clips with a 2-level 46-category genre hierarchy, indicating that – though the problem is clearly challenging – good category suggestions can be achieved. For example, if TubeFiler suggests 5 categories, it hits the right one (or at least its supercategory) in 91.8% of cases.

1. INTRODUCTION

Currently, web video services rely on a retrieval based on tags and keyword search. This approach makes only limited use of category information – for example, YouTube users can only choose between 13 categories, which are hidden in the "advanced search" options. In contrast to this, taxonomies are powerful tools that support browsing in other application areas (e.g., for document collections).

As a key requirement to make better use of taxonomies, we identify a richer (and particularly deeper) genre hierarchy. For example, it might be helpful to refine the category *sports* further into *sports/basketball*, *sports/hockey*, etc. However, while users would benefit significantly from such detailed categorization, their willingness to manually sort their videos into a deep taxonomy (and thus create an appropriate index) is usually limited.

To overcome this limitation, we present a system called *TubeFiler* that supports richer genre hierarchies by offering two key features:

1. **Automatic Categorization**: The core of our system is a categorizer, which inserts videos into a genre hierarchy using a statistical classification based on tags, titles, and visual features. This categorization can work fully automatically or in a semi-automatic fashion (e.g., by suggesting categories for a user's clip).
2. **Deep-level Clustering**: Due to the enormous amount and diversity of web video, the utility of any pre-defined taxonomy is limited. Therefore, to expand the hierarchy to even deeper levels, TubeFiler partitions genres further by *unsupervised* learning.

A web demo of our system is available at `http://www.dfki.uni-kl.de/~ulges/tubefiler`. Here, TubeFiler was used for a fully automatic categorization of ca. 1,000 test videos. The results of this categorization process can be browsed by clicking on the category name. Using deep-level clustering, TubeFiler further identifies groups of visually similar content within a category and displays them in a "cluster view" (see Fig. 1 for an illustration).

2. PROTOTYPE

TubeFiler performs an automatic multi-level categorization of YouTube clips (previous work has only tackled a single level so far [5]). On the two highest levels, videos are assigned to genres such as *movies*, *travel/city*, or *show/comedy* using supervised machine learning techniques and multimodal features. Results of different modalities are combined using a weighted sum fusion. These categories can be refined further by deep-level clustering, which identifies clusters of visually similar content.

Categorization by Tags Videos in YouTube are usually associated with titles or tags, which contain valuable semantic clues about a video's content. Hence, our approach puts strong weight on classification from such meta-information. For each category, a two-class linear kernel SVM classifier [3] is applied to "bag-of-word" features. These are normalized and scaled by each word's *inverse category frequency*, as words occurring in all categories are too unspecific. This approach proved to be superior over other combinations with RBF kernels or unweighted features in previous tests.

Visual Features Visual categorization is based on the well-known "bag-of-visual-words" model [4]. Clips are represented by keyframes, from which SIFT features are extracted and matched with a visual codebook of 4,000 clusters. The resulting features are fed to category-specific binary SVMs. Finally, SVM scores (mapped to class posteriors) are averaged over all keyframes.

*This paper addresses the Google Challenge: "Video Genre Classification"

†{damian.borth, joern.hees, markus.koch, adrian.ulges, christian.schulze, tmb}@dfki.de, rparedes@dsic.upv.es

ACM Multimedia '09 Beijing, China

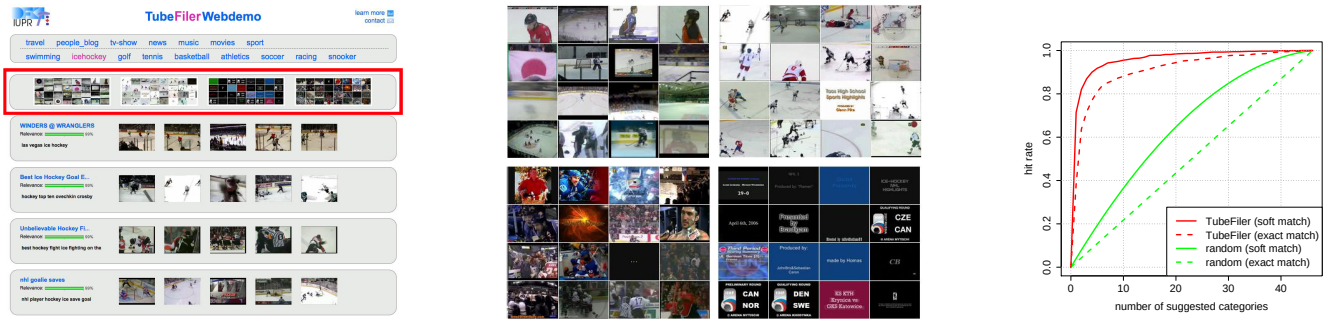

Figure 1: Left: By clicking on a category (here, *sport/hockey*), a user accesses all videos that TubeFiler has placed within. A "cluster view" (red box) also offers a deeper grouping into visually similar clusters. **Center**: Sample clusters found in the category *hockey* show "field shots", "crowd shots", or "title pages". **Right**: Our experimental results show that TubeFiler achieves a hit rate of 91.8% when suggesting 5 categories per clip.

Deep-level Clustering We refine genres further using a clustering of videos within a category. For this purpose, we employ Probabilistic Latent Semantic Analysis (PLSA) [2], which has been developed in the text domain to identify latent *topics* in document collections. We apply PLSA to *visual words*, obtaining clusters of visually similar content. The approach might also be applied to tags and titles.

3. EXPERIMENTS

Genres We have defined a two-level, 46-category genre hierarchy. Seven first-level categories like *news* and *sports* are refined into 39 more precise ones, like *music/hiphop* or *travel/beach*. Categories were chosen manually to reflect web video content well, including classical video genres like *movies/horror* as well as others specific to web content, like *people/videoblog*. Our annotations on $1,000$ randomly downloaded YouTube clips show that this hierarchy covers about 60% of YouTube content.

Dataset For each category, clips were downloaded from YouTube using manually formulated queries, such as (for *travel/city*) "sightseeing+city" or "trip to city". Also, downloads were restricted to a YouTube category. The resulting material was manually refined, obtaining 100 clips per genre category. To avoid a bias in the evaluation, the tags used for downloading a clip (here: "city", "sightseeing", and "trip") were ignored. For the same reason, training and test data were split by upload time, simulating a system trained in the past (more precisely, before Dec 15th 2008) applied to future data. This resulted in a training set of $3,502$ clips and a test set of $1,098$ clips.

Results For each test clip, TubeFiler suggest the N best categories. If the correct one was among the suggestions, we count the clip as a hit (*exact match*). We also evaluate a *soft match*, in which supercategories (*sports*) count as additional hits for subcategories (*sports/soccer*). Fig. 1 plots the hit rate for both measures vs. the number of suggestions N. As a baseline, a system based on random guessing was used.

TubeFiler performs best when combining textual and visual information (fusion weights were 0.6/0.4). Fig. 1 shows that, when suggesting 5 categories, the system gives a hit rate of 91.8%. It hits the *exact* category in 80.9% of cases. We also evaluated the benefits of different modalities (using soft matching). The tag-based approach gives 90.6%, which clearly outperforms a purely visual categorization (45.5%).

Finally, we also illustrate the results of deep-level clustering. Fig. 1 shows a result from the category *sports/hockey*. The system detects groups like "field shots", "crowd shots", and "title frames". Though for many categories results are not as impressive as in this example, our impression is that deep-level clustering provides a useful clue for browsing into categories in detail.

4. DISCUSSION

In this paper, we have presented TubeFiler, a system that automatically categorizes YouTube clips into a genre hierarchy using machine learning techniques and multi-modal features. We have demonstrated promising results, with a hit rate of 91.8% when suggesting 5 categories on a 46 class genre hierarchy. Thereby, tag information proved to be of vital importance, while visual information was found useful for an unsupervised clustering on deeper hierarchy levels.

The current prototype runs at sufficient speed for our lab tests: applying the classifier based on tags is highly efficient. For the visual modality, an additional feature extraction is required that is done in ca. 4.8 seconds per clip using a graphic hardware implementation[1]. It should be noted, however, that the approach is highly parallelizable on category level, on keyframe level, or even on the level of SVM kernel evaluations. Finally, for a scalable training, efficient on-line learning strategies might be investigated (e.g., [1]).[2]

5. REFERENCES

[1] D. Grangier and S. Bengio. A Discriminative Kernel-based Model to Rank Images from Text Queries. *IEEE PAMI*, 30(8):1371–1384, 2008.

[2] T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196, 2001.

[3] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[4] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proc. Int. Conf. Computer Vision*, pages 1470–1477, Oct. 2003.

[5] L. Yang, J. Liu, X. Yang, and X.-S. Hua. Multi-modality Web Video Categorization. In *Proc. MIR*, pages 265–274, 2007.

[1] `http://cs.unc.edu/~ccwu`

[2] This work was supported by the German Research Foundation (DFG), project MOONVID (BR 2517/1-1).