DFKI and University of Kaiserslautern Participation at TRECVID 2010 - Semantic Indexing Task

Damian Borth
Department of Computer Science, University of Kaiserslautern
d_borth@cs.uni-kl.de

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

Markus Koch
Department of Computer Science, University of Kaiserslautern
m_koch@cs.uni-kl.de

Thomas M. Breuel
Department of Computer Science, University of Kaiserslautern
tmb@cs.uni-kl.de

Abstract

Run No.	Run ID	Run Description	infMAP (%)
training on IACC data			
1	F_A_DFKI-MADM_3	SIFT visual words, Color Correlograms and Face-Detection separately trained, late fusion of SVMs scores	5.0
2	F_A_DFKI-MADM_4	SIFT visual words with SVMs	4.4
combined training on YouTube			
3	F_D_DFKI-MADM_1	SIFT visual words, Color Correlograms and Face-Detection separately trained, late fusion of SVMs scores	2.1
4	F_B_DFKI-MADM_2	SIFT visual words with SVMs	1.3

This paper describes the TRECVID 2010 participation of the DFKI-MADM team in the semantic indexing task. This years participation was dominated by two aspects, a new dataset and a large-sized vocabulary of 130 concepts. For the annual TRECVID benchmark this means to scale label annotation efforts to significant larger concept vocabularies and datasets.

Aiming to reduce this effort, our intention is to automatically acquire training data from online video portals like YouTube and to use tags, associated with each video, as concept labels. Results for the evaluated subset of concepts show similarly to last year's participation [3], that effects like label noise and domain change lead to a performance loss (infMAP 2.1% and 1.3%) as compared to purely TRECVID trained concept detectors (infMAP 5.0% and 4.4%). Nevertheless, for individual concepts like "demonstration_or_protest" or "bus" automatic learning from online video portals is a valid alternative to expected labeled training datasets. Furthermore, the results also show that fusion of multiple features helps to improves detection precision.

1. Introduction

As video databases are growing in size [9] the demand for robust search and retrieval tools increases. One successful strategy for such tools is concept-based video retrieval [15], which can be splitted into *concept detection* building a semantic index and *video search* using such an index.

The semantic indexing task of this year's TRECVID benchmark deals with (1) an increase of concept vocabulary size to 130 concepts and (2) a new dataset consisting of videos from the Internet Archive. Due to the supervised learning approach used in many systems, this leads to the demand of acquiring labels for each of the 130 concepts - a very time-consuming and costly effort. As in previous years, the TRECVID community copes with this by a collaborative annotation effort [2], which this year - however - was expected to demand much more man-hours of annotation work than during the previous switch of datasets [20].

Recently, socially tagged images and video have been used as training sources for semantic indexing [11, 19]. Such data is publicly available at large scale from online portals like Flickr or YouTube and is associated with a noisy but rich corpus of tags, comments and ratings that are provided by their online communities. Being able to automatically learn new concepts from such online sources can reduce the need for large scale acquisition of expert labels and seamlessly increase concept vocabularies resulting in retrieval systems being scalable w.r.t. user's information need [6].

On the downside, the usage of web videos as training material for concept detection systems faces new challenges, as its utility as training data strongly dependents on user generated tags. For example, tagged video clips are often subjectively annotated, which leads to unreliable and coarse labels (only a fraction between 20%-50% of web video is relevant as estimated in [16]). Second, in a setup where concept detectors are trained on user generated video content and afterwards applied to a different domains like the "IACC" dataset used in TRECVID, we are facing the so-called domain change problem: a significant discrepancy of the visual characteristics between different video sources. Both effects are known to cause a significant performance loss of concept detection systems [10, 16].

Table 1. Queries for Training Set Acquisition from YouTube for the evaluated concepts

concept	YouTube query	YouTube category
Airplane flying	airplane flying -indoor -school	Autos&Vehicles
Animal	animal dog cat horse birds	-
Asian People	aisan -hot -sexy -bikini	People&Blog & Entertainment
Bicycling	riding bicycle fahrrad	Sports
Boat_Ship	ship queen freedom royal	Autos&Vehicles
Bus	bus -van -suv -vw -ride	Autos&Vehicles
Car Racing	car racing -rc	Sports
Cheering	classroom cheering applauding	Entertainment
Cityscape	cityscape -slideshow	Travel&Places
Classroom	classroom school -secret	-
Dancing	people dancing learn to dance	Sports
Dark-skinned People	black people	-
Dem._Or_Prot.	protesting	-
Doorway	türen öffnen doors gates	People&Blog & Entertainment
Explosion_Fire	explosion	How-To&DIY
Female-human-face-closeup	female videoblog girl makeup	People&Blog & HowTo&Style
Flowers	flower bouquet bloom	-
Ground_Vehicle	car bus tank emergency vehicle truck car racing	-
Hand	hand daft	-
Mountain	mountain panorama	Travel&Places
Nighttime	by night	Travel&Places
Old_People	old people	-
Running	running athletics	Sports
Singing	singing gospel choir	-
Sitting_Down	sitting down restaurant scene table	-
Swimming	swimming	Sports
Telephones	phone & device	-
Throwing	throwing -potery -cement	Sports
Vehicle	car bus tank emergenvy vehicle truck car racing	-
Walking	walking people -running	Travel&Places

2. Datasets

Two different datasets were used for system training: first, a collection of video clips downloaded from YouTube (referred to as YOUTUBE), which provides user generated tags being used as concept labels. This dataset is used as retrieved from YouTube i.e. no manual filtering or further processing was done. The second dataset is the IACC.1 data (referred to as TRECVID), which is a new dataset introduced in the TRECVID 2010 benchmark. For this dataset concept labels have been acquired by manual inspection through TRECVID's collaborative annotation effort.

The download of YouTube videos was performed in two steps. First, the YouTube API[1] was used to retrieve meta-data of potential video clips. This was done by manually mapping a concept definition to a textual query like "mountain landscape - biking" embedded in the API call. Such a mapping must be done carefully to prevent concept drifts i.e to narrow down retrieval to videos matching the concept definition given by TRECVID as closely as possible. The manual query mapping was performed on two different levels (a complete list of final queries for the selected 30 concepts is given in Table 1):

1. YouTube is organizing videos in categories like "Sport" or "Autos&Vehicles". For some concepts, we enhanced the query with a canonical category, which restricted the list of retrieved videos to this category. For example, by choosing the category "People&Blog" we could improve the quality of video material for the concept "female human face closeup" in getting more video clips of closeup faces.

2. Queries were additionally refined by inspecting of YouTube search results and accordingly adding or excluding additional keywords. For example, for the concept "ground vehicle" we added the keyword "car" or for the concept "cityscape" we excluded the term "slideshow".

After defining queries and retrieving meta-data of potential video clips, we downloaded 150 videos for each new concept from YouTube. To reduce data load we only downloaded the first 3 minutes of each clip, resulting in a training set of about 19,000 videos. For a subset of 53 concepts we were using video material TubeTagger, a system which performs visual learning on YouTube clips [17].

[1] http://code.google.com/apis/youtube/overview.html

TODO Figure 1 is displaying random sample keyframe from both training sets *YOUTUBE* and *TRECVID* for the representative concepts "person_playing_soccer", "traffic_intersection", "person_eating" and "female_human_face_closeup". It can be seen that while YOUTUBE grasps the concept definition it contains a significant amount of non-relevant content. Note that this non-relevant material will also be used as positive samples in our concept detector training.

3. Approach

In this year's TRECVID participation, we used a standard concept detection pipeline consisting of SIFT visual word features and SVMs classifier. Additionally to last year's participation, color correlograms and face detection features have been evaluated and the focus lies entirely on SVM classifiers. The system is describes as following:

3.1 Keyframe Extraction

Regarding shot representation we extract keyframes for each video/shot. Here, we deal differently with the given datasets:

For the *YOUTUBE* data, keyframe extraction is performed according to a change detection scheme [17] providing 125,000 keyframes for the entire dataset, which corresponds to an average of ca. 7 keyframes per YouTube video clip.

For the *TRECVID* data, the standard shot boundary reference was used for temporal segmentation and an intra-shot diversity based approach for keyframe extraction [4]. For each shot, a K-Means clustering is performed over MPEG7 Color Layout Descriptors [13] extracted from all frames. The number of clusters is fitted using the Bayesian Information Criterion [14]. For each cluster the frame closest to the cluster center is chosen as a keyframe.

3.2 Features

For each keyframe the following visual features are extracted:

- **Visual Words (SIFT)**: Visual words are extracted by performing a dense regular sampling of SIFT features [12] at several scales, obtaining ca. 3,600 features per keyframe. Features are clustered to 5,000 visual words using K-Means, obtaining a "bag-of-visual-words" descriptors.

Figure 1. TODO: some mosaic imgs from either randomly selected keyframes or top ranked result keyframes.

- **Color Correlograms**: To capture color information, color correlograms [8] have been extracted. The descriptor forms a 600-dimensional vector and is normalized to 1 [7].
- **Face Detection**: We employed OpenCV's standard frontal face detector[2] as a basis for this third feature. The number and average size of faces detected in the image were combined to a two-dimensional feature vector. Thereby, the size was normalized to mean 1 and standard deviation 0.75 (if no face was found, this feature was set to -1).

3.3 Statistical Model

Support vector machines (SVMs) were used as a standard approach for concept detection, forming the core of numerous concept detection systems [15]. We used the LIBSVM [5] implementation with a χ^2 kernel, which has empirically been demonstrated to be a good choice for histogram features [21]:

$$K(x,y) = e^{-\frac{d_{\chi^2}(x,y)^2}{\gamma^2}} \quad (1)$$

where $d_{\chi^2}(.,.)$ is the χ^2 distance. γ and the SVM cost of misclassifications C were estimated separately for each concept using a grid search over the 3-fold cross-validated average precision. A problem is that training sets are *imbalanced*, i.e. the number of negative samples outnumbers the number of positive ones. Those setups cause problems for many classifiers, including SVMs [1]. To overcome this problem, the dominant class is subsampled to obtain roughly balanced training sets. For the TRECVID based runs, SVMs were trained on the given small-scale training sets, and the results were fused using a simple averaging. For the YouTube-based runs (where significantly more positive training samples were available), we used 3000 positive and 6000 negative training examples from the *YOUTUBE* data set.

In all cases, SVM scores were mapped to probability estimates using the LIBSVM standard implementation.

[2] http://opencv.willowgarage.com/wiki/FaceDetection

3.4 Black Frames Removal

The majority of YouTube's database consist of user generated content. Such video clips often contain intro and outro frames with text overlays. Following, the trained detectors are sensitive to such frames resulting in false positives. To prevent such behavior we post-processed the final result lists and filtered high ranked keyframes which were mostly or entirely black.

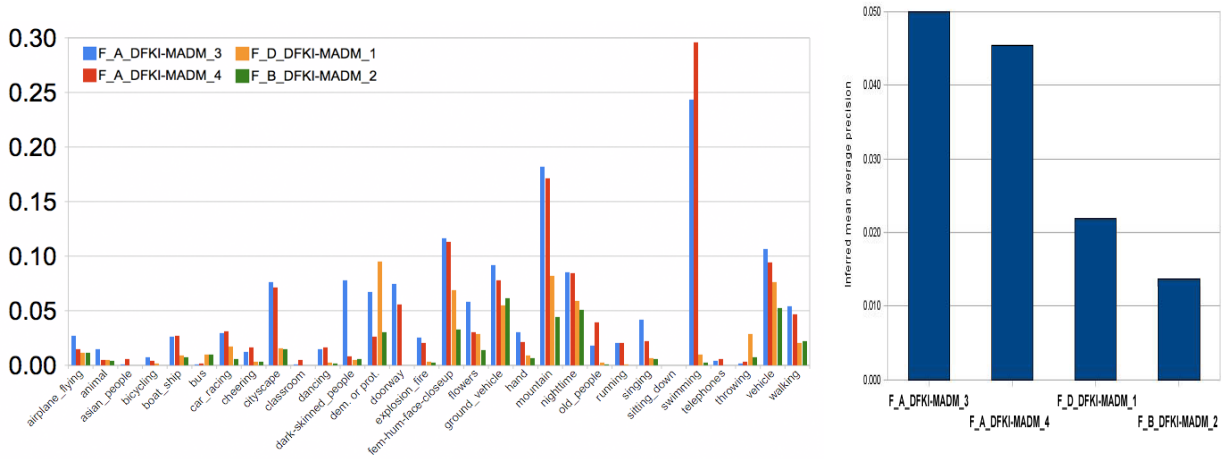

Figure 2. Quantitative results for all runs. The first two runs are using TRECVID data for training the last two ones display results from detectors trained on YOUTUBE data. left**: per-concept results.** right**: the mean inferred average precision per run.**

3.5 Late Fusion

Finally, scores obtained from several keyframes are fused:

- Having several keyframes for each shot, the corresponding scores are simply averaged, providing a single score for each shot and feature.
- For fusing different features, we perform a weighted sum fusion whereas concept-specific weights are learned using a grid search maximizing average precision on the TRECVID data set.

4 Results

We submitted 4 runs for the full submission including all 130 concept detections. In particular, 2 runs have been trained on *TRECVID* data and 2 runs on *YOUTUBE* data:

1. **F_A_DFKI-MADM_3** In this run, we used the SVM approach in combination with SIFT visual word features, color correlograms and face detection trained on TRECVID data.
2. **F_A_DFKI-MADM_4** As in F_A_DFKI-MADM_3, the second run used SVMs only with SIFT visual word features being trained on TRECVID data. It illustrated the benefit of multiple features fusion when compared pure SIFT based visual word features.
3. **F_D_DFKI-MADM_1** In contrast to the F_A_DFKI-MADM_3 setup, in this run we trained the detectors on YOUTUBE material.
4. **F_B_DFKI-MADM_2** Here, we perform concept detection as described in F_A_DFKI-MADM_4 but training is done entirely on YOUTUBE data.

Quantitative results are displayed in Figure 2. It can be seen that concept detection using multiple features (infMAP of 5.0%, 2.1% for F_A_DFKI-MADM_3 and F_D_DFKI-MADM_1) outperforms pure SIFT visual word concept detection (infMAP of 4.4%, 1.3% for F_A_DFKI-MADM_4 and F_B_DFKI-MADM_2). Also, as being observed in previous TRECVID benchmarks [3, 18] a significant domain change leads to a performance loss being quantified by 2.9% when comparing TRECVID trained detectors against YOUTUBE trained ones.

5 Discussion

The finding of this year's TRECVID participation are ... - domain change - duplicates - redundancy in the data - easy data acquisition - need just 6 hours of manual intervention to acquire the 130 concepts

6 Acknowledgements

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG), project MOONVID (BR 2517/1-1).

References

[1] R. Akbani, S. Kwek, and N. Japkowicz. Applying Support Vector Machines to Imbalanced Datasets. In *Proc. Europ. Conf. Machine Learning*, pages 39–50, 2004.

[2] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In *Proc. Europ. Conf. on Information Retrieval*, pages 187–198, 2008.

[3] D. Borth, M. Koch, A. Ulges, M. Koch, and T. Breuel. DFKI-IUPR Participation in TRECVID'09 High-level Feature Extraction Task. In *Proc. TRECVID Workshop*, November 2009.

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

[5] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001.

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

[7] J. Hofmann and M. Ali. An extensive approach to content based image retrieval using low- and high-level descriptors. Diploma Thesis, University of Gteborg, Sweden, 2006.

[8] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu, and R. Zabih. Image Indexing Using Color Correlograms. In *Proc. Int. Conf. on Pattern Recognition*, page 762, 1997.

[9] R. Junee. Zoinks! 20 Hours of Video Uploaded Every Minute! The YouTube Blog; available from `http://www.youtube.com/blog?entry=on4EmafA5MA` (retrieved: May'09), May 2009.

[10] X. Li, C. Snoek, and M. Worring. Learning Tag Relevance by Neighbor Voting for Social Image Retrieval. In *Proc. Int. Conf. on Multimedia Information Retrieval*, pages 180–187, October 2008.

[11] X. Li, C. Snoek, and M. Worring. Annotating Images by Harnessing Worldwide User-Tagged Photos. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, April 2009.

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

[13] B. Manjunath, J.-R. Ohm, V. Vasuvedan, and A. Yamada. Color and Texture Descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[14] G. Schwarz. Estimating the Dimension of a Model. *Ann. of Stat.*, 2(6):461–464, 1978.

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

[16] A. Ulges, D. Borth, and T. Breuel. Visual Concept Learning from Weakly Labeled Web Videos. In *Video Search and Mining*. Springer-Verlag, 2009.

[17] A. Ulges, M. Koch, D. Borth, and T. Breuel. TubeTagger YouTube-based Concept Detection. In *Proc. Int. Workshop on Internet Multimedia Mining*, December 2009.

[18] A. Ulges, M. Koch, C. Schulze, and T. Breuel. Learning TRECVID'08 High-level Features from YouTube™. In *Proc. TRECVID Workshop*, November 2008.

[19] A. Ulges, C. Schulze, M. Koch, and T. Breuel. Learning Automatic Concept Detectors from Online Video. *Comp. Vis. Img. Underst. (accepted for publication)*, 2009.

[20] J. Yang and A.G. Hauptmann. A framework for classifier adaptation and its applications in concept detection. 2008.

[21] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *Int. J. Comput. Vis.*, 73(2):213–238, 2007.