Learning TRECVID'08 High-Level Features from YouTube™

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

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

Christian Schulze
German Research Center for Artificial Intelligence (DFKI)
christian.schulze@iupr.dfki.de

Thomas M. Breuel
TU Kaiserslautern and DFKI,
breuel@iupr.dfki.de

Abstract

Run No.	Run ID	Run Description	infMAP (%)
training on TV08 data			
1	IUPR-TV-M	SIFT visual words with maximum entropy	6.1
2	IUPR-TV-MF	SIFT with maxent, fused with color+texture, and motion (NN matching)	5.9
3	IUPR-TV-S	SIFT visual words with SVMs	5.3
4	IUPR-TV-SF	SIFT with SVMs, fused with color+texture, and motion (NN matching)	6.3
training on YouTube™data (no use of standard training sets)			
5	IUPR-YOUTUBE-S	SIFT visual words with SVMs	2.2
6	IUPR-YOUTUBE-M	SIFT visual words with maximum entropy	2.1

We participated in TRECVID's High-level Features task to investigate online video as an alternative data source for concept detector training. Such video material is publicly available in large quantities from video portals like YouTube™. In our setup, tags provided by users during upload serve as weak ground truth labels, and training can scale up to thousands of concepts without manual annotation effort. On the downside, online video as a domain is complex, and the labels associated with it are coarse and unreliable, such that performance loss can be expected compared to high-quality standard training sets.

To find out if it is possible to train concept detectors on online video, our TRECVID experiments compare the same state-of-the-art (visual only) concept detection systems when (1) training on the standard TRECVID development data and (2) training on clips downloaded from YouTube™. Our key observation is that youtube-based detectors work well for some concepts, but are overall significantly outperformed by the "specialized" systems trained on standard TRECVID'08 data (giving a infMAP of 2.2% and 2.1% compared to 5.3% and 6.1%). An in-depth analysis of the results shows that a major reason for this seems to be redundancy in the TV08 dataset.

1. Introduction

While the automatic detection of semantic concepts (or "high-level features") in video streams is a key component of research prototypes for content-based video search, a critical burden for its practical application is that the underlying machine learning techniques require annotated training sets. Since target concepts can be visually complex, hundreds of sample shots per concepts may be needed. Also, the number of concepts required to cover users' information needs is high.

For standard training sets, annotations are acquired manually and explicitly for the purpose of concept detector training. Since this is a time-consuming (and thus cost-intensive) process, researchers share annotations [13, 17] or organize collaborative labelling efforts [2]. This has made it possible to train detectors for several hundred concepts (e.g., [23]). Yet, several problems remain with the explicit acquisition of ground truth:

1. One obvious problem is that – to cover users' information needs – the number of concepts to be trained must be increased by a further magnitude compared to the state-of-the-art (current estimates for a sufficient number of concepts are in the range of $3.000 - 5.000$ [8]).

2. Current detectors are mostly trained on a single annotated video collection (often the TRECVID'05 dataset). The resulting detectors work well on this dataset (or very similar ones) but generalize poorly, as has been demonstrated in [24].

3. Manual annotations are static, and so are the concept detectors trained on them. In contrast to this, the world around us – and with it its videos and users' information needs – is constantly evolving. New concepts of interest emerge, like "9-11", "secondlife", or "Barak Obama". Similarly, concept detection systems should adapt to dynamic user interest, which is impractical using explicit manual annotations.

To overcome these problems to some extent, we propose to investigate an alternative data source, namely *online video* that is publicly available at a large scale from portals like YouTube$^{\text{TM}}$, blinkx, and many others. These web videos are enriched with textual descriptions that can serve as weak annotations in a machine learning framework for training concept detectors. This way, manual annotation effort is shifted to the youtube community, and a concept detection system can learn autonomously by acquiring its readily annotated training set from the web. This setup offers the advantages of *scalability* (it is possible to scale concept detectors up to several thousands of concepts) and *flexibility* (web video content is kept up-to-date by the community, such that concept detectors trained on it can keep track of concepts that change or emerge).

On the downside, the labeling information that comes with web video clips is of a significantly lower quality than that of manually annotated datasets currently used. This is due to several reasons: first, while in TRECVID videos are labeled on shot level, youtube tags are given on video level (and not all shots might be visually related to a tag). Second, shots are usually annotated according to clear visual criteria, like "shots that take place outdoors at night, but no sporting events under lights" (LSCOM concept no. 352), tags at web video portals are often given with an intention that links the tag only indirectly to the visual content. Consequently, the training sets acquired from YouTube$^{\text{TM}}$contain relevant material as well as "junk" frames not visually related to the target concept. The key question arising from this fact is: Can concept detectors successfully be trained on online video?

To give an answer, we participated in TRECVID's High-level Features task and present our experiences with training a concept detection system for TRECVID'08 on YouTube$^{\text{TM}}$. For a quantitative evaluation, our general strategy is to train a state-of-the-art concept detection approach (namely discriminative training over bag-of-visual-words features) on two different data sources: (1) the standard training set of TRECVID'08 (referred to as `TRECVID`in the following), and (2) a set of tagged videos downloaded from youtube (called `YOUTUBE`). We first describe both datasets (particularly, the acquisition of the `YOUTUBE`dataset). After this, the concept detection approach is briefly discussed, and experimental results are provided. Finally, a discussion of our results is given.

2. Datasets

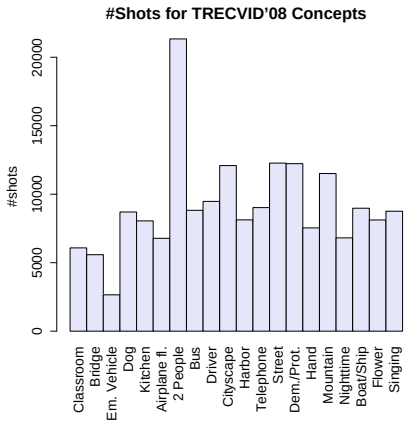
To investigate how well a state-of-the-art concept detector performs when trained on online videos, we compare the same approach for two sources of training data: first, the standard TRECVID'08 development data with annotations provided by the Chinese Academy of Sciences

Table 1. Queries for Training Set Acquisition from youtube.

concept	youtube query	youtube category
Classroom	classroom & school -secret	-
Bridge	bridge -crossing -ship	Travel&Places
Em._Vehicle	emergency & vehicle -driver -ride	Autos&Vehicles
Dog	dog	Pets&Animals
Kitchen	kitchen -knife -remodel	Howto&Style
Airplane_flying	airplane & flying -jefferson -indoor -school -kids	Autos&Vehicles
Two_people	two & people -sleepy -questions	People&Blogs
Bus	bus -van -suv -vw -ride	Autos&Vehicles
Driver	car & vehicle & driver -simulator	Autos&Vehicles
Cityscape	cityscpae -slideshow -emakina	Travel&Places
Harbor	harbor & industry & ship	-
Telephone	phone & device	-
Street	street & paved	-
Dem._Or_Prot.	protesting	-
Hand	hand & daft	-
Mountain	mountain & panorama	Travel&Places
Nighttime	by & night	Travel&Places
Boat_Ship	ship & (queen | freedom | royal)	Autos&Vehicles
Flower	flower & (bouquet | bloom)	-
Singing	singing & (gospel | choire)	-

(TRECVID). Second, a dataset of video clips was downloaded from YouTube™, whereas video-level annotations for training are taken from video descriptions provided by youtube users during upload (YOUTUBE).

A first interesting question is whether a sufficient *quantity* of training data can be obtained from YouTube™. While the portal offers a tremendous overall amount of video data (83.4 Mio. clips by April 2008 [25]), it is not clear a priori how much training material is actually available for typical target concepts, since the distribution of concepts is highly biased towards popular tags (e.g., "funny", "love", or "girl"). To investigate how much material is available for standard concepts, we downloaded meta-data for up to 1000 video clips per concept (this upper limit is imposed by youtube). From these clips, the number of shots obtained per concept was estimated by assuming 4.79 shots per minute. Figure 1 plots the result for



(a)

Figure 1. Estimated quantity of training shots obtainable from YouTube™for the 20 TRECVID'08 concepts.

the 20 concepts used in the TRECVID'08 evaluation. It can be seen that a fair number of shots can be obtained for most concepts (9146 on average). This quantity is significantly higher than for the TRECVID'08 standard set (avg. 481 annotations per concept). Two outliers can be observed: for the concept "emergency_vehicle", only 2654 shots are estimated. For the concept "two_people", 21337 shots are obtained.

For our concept detection experiments, clips from youtube were acquired by simulating queries to the youtube API[1]. To improve the quality of the downloaded material, two refinements were done manually (the exact list of final queries is given in Table 1):

1. Videos at YouTube™are organized in categories like "Pets&Animals" or "Autos&Vehicles". For some concepts, a canonical category was picked, and the download was restricted to this category. For example, by restricting the download for "bridge" to "Travel&Places", the music video "Bridge over troubled Water" by Simon&Garfunkel is excluded from the training set.

2. The queries were refined according to a brief qualitative analysis of youtube search results. For example, for "mountain" the term "panorama" was added, or for "kitchen" the term "knife" was excluded.

[1] http://www.youtube.com/dev

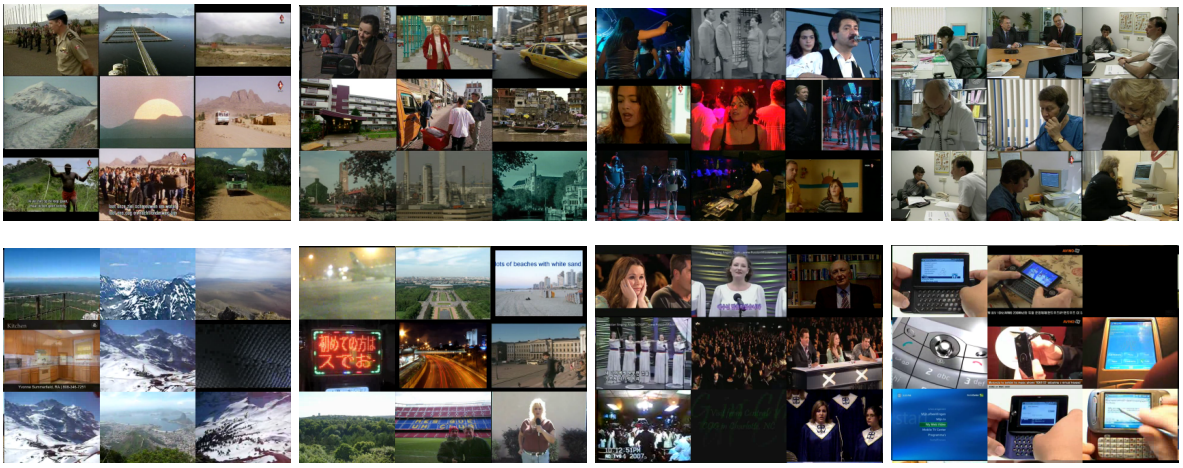


Figure 2. An illustration of randomly selected key frames from TRECVID(top) and YOUTUBE(bottom) for the concepts "mountain", "cityscape", "singing", and "telephone". While the TRECVID dataset shows high annotation quality, the material downloaded from youtube contains a significant amount of junk.

To reduce the data load for training, we only downloaded 100 videos per concept of up to 3 minutes length. This gave a training set of 42 hours. Sample keyframes for both training sets (YOUTUBEand TRECVID) are illustrated in Figure 2. Generally, it can be seen that the YOUTUBEdata contains relevant material together with significant amounts of junk. Also, a difference between concepts can be observed: our impression is that training from YouTube^{TM}works best for concepts that youtube users find interesting enough to film, edit, and upload. For example, for the concept "mountain" the YOUTUBEdataset contains lots of panoramic views that make good training samples, while for other concepts like "cityscape" lots of junk frames can be found. For the concept "telephone", youtube videos tagged with the concept tend to show close-ups of phones, whereas TRECVIDshots show people telephoning. Here, though the concept is represented well in both training sets, the appearance differs due to a *domain change* between both sets. Overall, however, it seems reasonable to assume that at least some concepts can successfully be learned from youtube material.

3. Approach

The purpose of our TRECVID experiments is to evaluate the same state-of-the-art concept detection system when training on the standard TRECVID development data and training on clips downloaded from YouTube^{TM}. Thereby, the core of the concept detection system used is a standard approach (SIFT visual words + discriminative training). This approach has been demonstrated to work well in several visual recognition tasks [7, 15, 9, 6], including concept detection [21]. Further, in Runs 2 and 4, a weighted sum fusion with other feature modalities (color+texture and motion) is used. Details are outlined in the following.

3.1 Keyframe Extraction

Instead of using only a single keyframe per shot, we capture intra-shot diversity due to scene changes and camera motion using an adaptive two-step approach (for further information, see [3]):

1. For the YOUTUBEdata, shot boundary detection is performed using an adaptive thresholding over color descriptor differences [10]. For the TRECVID data, the standard shot boundary reference was used.

2. Within each shot, a K-Means clustering is performed over MPEG7 Color Layout Descriptors [12] extracted from all frames. For each cluster, the frame closest to the center is extracted as a keyframe. The number of clusters is determined using the Bayesian Information Criterion (BIC) [16], which balances the number of keyframes explaining the shot versus the fitting error.

Using this method, we obtain ca. 3 keyframes on average per shot, which corresponds to an overall of 35.943 keyframes for the YOUTUBEtraining set, 112.867 for the TRECVIDtraining set, and 112.301 for the TRECVID'08 test set.

3.2 Features

From all keyframes the following visual features are extracted:

1. **Visual Words (SIFT)**: Visual words are extracted by performing a dense regular sampling of SIFT features [11] at several scales, obtaining ca. 3600 features per keyframe. Features are clustered to 2000 visual words using K-Means. The resulting "bag-of-visual-words" descriptors are combined with discriminative SVM and maximum entropy classifiers (see Section 3.3), forming the core of all submitted runs.

2. **Color+Texture**: Optionally, a combination with other feature modalities can be included. For this purposes, simple descriptors for color ($8 \times 8 \times 8$ RGB histograms) and texture (histograms over Tamura features [18]) were extracted and concatenated in an early fusion. These features are combined with nearest neighbor matching (see Section 3.3)

3. **Motion**: To capture discriminative motion patterns, tiled histograms over MPEG-4 motion are extracted using the codec XViD[2]. Like color+texture descriptors, these features are combined with nearest neighbor matching (see Section 3.3). For more details, please refer to our previous publications [19, 20].

3.3 Statistical Models

Three different statistical models are used:

- **Support Vector Machines**: As one statistical model, support vector machines (SVMs) are used. These are a standard approach for concept detection and are used in numerous systems [22, 23]. We used the LIBSVM [4] implementation with a χ^2 kernel, which has empirically been demonstrated to be a good choice for histogram features [26]:

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

where $d_{\chi^2}(.,.)$ is the χ^2 distance between histograms. 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 TRECVIDbased runs, 5 SVMs were trained on small-scale training sets with 400 negative samples randomly sampled from the TRECVIDset, 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 YOUTUBEdata set.

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

- **Maximum Entropy**: As an alternative to SVMs, we also test a different discriminative approach based on the maximum entropy principle, which has successfully been applied to object recognition before [5]. The posterior is modeled in a log-linear fashion:

$$P(t|x) \propto \exp\left(\alpha_t + \sum_{c=1}^{500} \lambda_{tc} h^c(x)\right), \quad (2)$$

where $h^c(x)$ is entry number c in the visual word histogram for frame x. The parameters $\{\alpha_t, \lambda_{tc}\}$ are estimated from a training set of tagged frames using an iterative scaling algorithm [5].

- **Nearest Neighbor Matching**: For the color+texture and motion features a nearest neighbor matching is used: given a training set of features representing tagged keyframes Y, we find the nearest neighbor $x' := \arg min_{y \in Y} ||y - x||_2$ for keyframe x, and the score for a tag t equals a vote for the tag of this neighbor (to realize fast nearest neighbor matching, an approximate search using a kd-tree is used [14]):

$$P(t|x) := \delta(t, t(x')) \quad (3)$$

[2] www.xvid.org

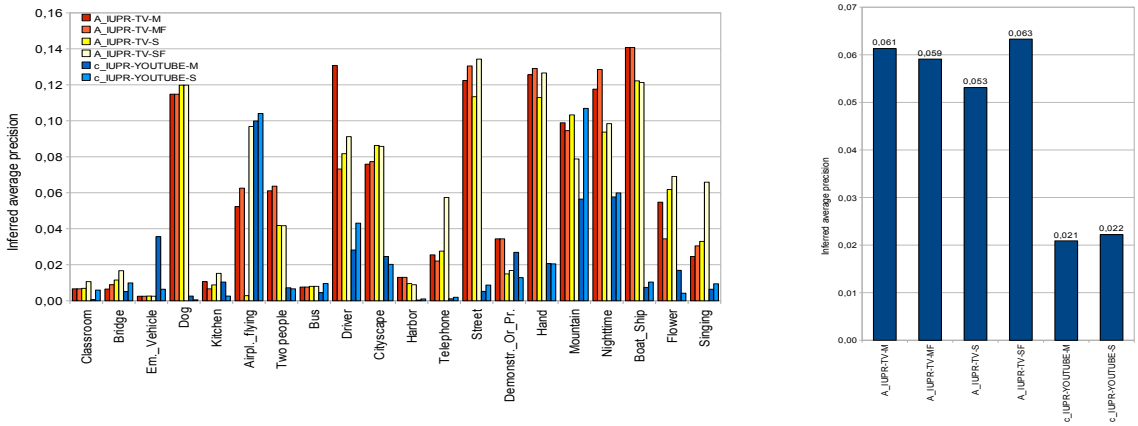


Figure 3. Quantitative results for all IUPR runs (the first four runs are trained on the TRECVID'08 standard data, the last two on youtube). Left: per-concept results. Right: the mean inferred average precision per run.

3.4 Late Fusion

Finally, scores obtained from several keyframes for each shot and feature scores for each keyframe (in Runs 2 and 4) are fused (no inter-concept fusion is done):

- 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 TRECVID2007 test set, using the TRECVID2007 devel set as training set. After re-training on the TRECVID2008 devel data, these weights are used to fuse the different features into a final concept score.

4 Results

We submitted a total of 6 runs: 4 runs trained on the TRECVIDdata, 2 runs trained on data obtained from YouTube™:

1. **A_IUPR-TV-M** In this run, we used the maximum entropy approach in combination with SIFT visual word features.
2. **A_IUPR-TV-MF** For this run, we fused the scores from Run 1 with results of nearest neighbor matching on color+texture and motion features.
3. **A_IUPR-TV-S** In contrast to Run 1, the maximum entropy model was replaced with SVMs.
4. **A_IUPR-TV-SF** The scores from Run 3 are fused with the results of NN matching on the color+texture and motion features.
5. **c_IUPR-YOUTUBE-S** Same as in Run 3 but using YOUTUBEdata as a training source.
6. **c_IUPR-YOUTUBE-M** Same as in Run 1 but using YOUTUBEdata as training source.

Quantitative results are illustrated in Figure 3. The youtube-based detectors perform comparable to standard detectors for a few concepts like "airplane" or "mountain", but are overall outperformed by the "specialized" systems trained on the standard training set, giving infMAPs of 2.1% (MaxEnt) and 2.2% (SVMs) as opposed to 6.1% and 5.3% given by the standard training sets. An in-depth look at the detections of the specialized detector reveals the reason for this: for "dog", the concept for which the difference between youtube and standard detectors is the most significant, detection results of a standard detector are illustrated in Figure 4. It can be seen that the TV08-based detector uses redundant material appearing in both training and testing and thus significantly outperforms the youtube-based detector. Obviously, for

Figure 4. Top: Detection results for the concept "dog" for a TV08-based detector. Bottom: "Dog" training samples in TV08. Obviously, the detector makes use of redundant material appearing in both training and testing, which is why it significantly outperforms the youtube-based detector.

the TV08 data similar findings hold as for the TV05 data, where 20% of the shots in are claimed to have duplicates in the training set [22]. This reveals that a major reason for the higher performance of the specialized detector is redundancy in the underlying video dataset.

Figure 5 illustrates top detection results of the youtube-based detector. For all concepts, it can be seen that the detectors are attracted by material similar to the training samples in Figure 2. For example, for "mountain" panoramic scenes are detected and detector performance can be considered good (quantitative results for this concept are comparable to the TV08 detector). For "telephone", the system is attracted by close-ups of devices and computer screens, which is similar to the training content but gives poor quantitative scores.

5 Discussion

Our key result from our participation in TRECVID'08's High-level Features task is that youtube-based detectors give reasonable detection results for some concepts, but are significantly outperformed by the "specialized" systems trained on the standard training set. We argue that the major reason for this is that standard detectors trained on TV08 make use of annotations in the target domain and exploit redundancy in the dataset.

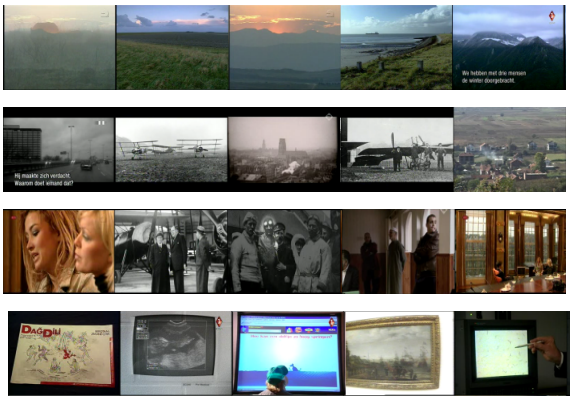


Figure 5. Top detections of the youtube-based detector for the concepts "mountain", "cityscape", "singing", and "telephone".

This raises the question how YouTube^{TM}videos compare to standard training sets if applying concept detectors to novel target domains unseen in training. We currently investigate this question in further experiments, as well as another issue, namely whether the generalization capabilities of detectors can be improved by enriching current standard training sets with material from YouTube^{TM}.

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, September 2004.

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

[3] Damian Borth, Adrian Ulges, Christian Schulze, and Thomas M. Breuel. Keyframe Extraction for Video Tagging and Summarization. In *Proc. Informatiktage 2008*, pages 45–48, 2008.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] T. Deselaers, D. Keysers, and H. Ney. Discriminative Training for Object Recognition Using Image Patches. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 157–162, June 2005.

[6] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-Visual-Words Models for Adult Image Classification and Filtering. In *Proc. Int. Conf. Pattern Recognition (accepted for publication)*, December 2008.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. Technical report, PASCAL Challenge Workshop. available from: `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html`, October 2007.

[8] A. Hauptmann, R. Yan, and W. Lin. How many High-Level Concepts will Fill the Semantic Gap in News Video Retrieval? In *Proc. Int. Conf. Image and Video Retrieval*, pages 627–634, Jul 2007.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proc. Int. Conf. Computer Vision and Pattern Recognition*, pages 2169–2178, June 2006.

[10] R. Lienhart. Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. *Int. J. of Img. and Graph.*, 1(3):469–286, 2001.

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

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

[13] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[14] R. Paredes and A. Perez-Cortes. Local Representations and a Direct Voting Scheme for Face Recognition. In *Proc. Workshop on Pattern Rec. and Inf. Systems*, pages 71–79, July 2001.

[15] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars. A Thousand Words in a Scene. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007.

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

[17] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *Proc. Int. Conf. Multimedia*, pages 225–226, October 2006.

[18] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Trans. System, Man, Cybernetics*, 8(6):460–472, 1978.

[19] Adrian Ulges, Christian Schulze, Daniel Keysers, and Thomas M. Breuel. Content-Based Video Tagging for Online Video Portals. In *Proc. MUSCLE/ImageCLEF Workshop*, September 2007.

[20] Adrian Ulges, Christian Schulze, Daniel Keysers, and Thomas M. Breuel. A System that Learns to Tag Videos by Watching Youtube. In *Proc. Int. Conf. on Vision Systems*, pages 415–424, May 2008.

[21] K. van de Sande, T. Gevers, and C. Snoek. A Comparison of Color Features for Visual Concept Classification. In *Proc. Int. Conf. Image and Video Retrieval*, pages 141–150, July 2008.

[22] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video Diver: Generic Video Indexing with Diverse Features. In *Proc. Int. Workshop Multimedia Information Retrieval*, pages 61–70, September 2007.

[23] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Technical report, Columbia University, March 2007.

[24] Jun Yang and Alexander G. Hauptmann. (Un)Reliability of video concept detection. In *Proc. Int. Conf. Image and Video Retrieval*, pages 85–94, July 2008.

[25] "Youtube". in Wikipedia: The Free Encyclopedia; (Wikimedia Foundation Inc.) [encyclopedia on-line]; available from `http://en.wikipedia.org/wiki/YouTube` (retrieved: Sep'08).

[26] 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.