SEMANTIC CLASS DETECTORS IN VIDEO GENRE RECOGNITION

Michal Hradiš [1], Ivo Řezníček[1] and Kamil Behúň[1]
[1]*Department of Computer Graphics and Multimedia, Brno University of Technology, Bozetechova 1/2, Brno, Czech Republic*
{*ihradis, ireznice*}*@fit.vutbr.cz, xbehun03@stud.fit.vutbr.cz*

Keywords: Video Genre Recognition, Semantic Indexing, Local Features, SIFT, SVM

Abstract: This paper presents our approach to video genre recognition which we developed for MediaEval 2011 evaluation. We treat the genre recognition task as a classification problem. We encode visual information in standard way using local features and Bag of Word representation. Audio channel is parameterized in similar way starting from its spectrogram. Further, we exploit available automatic speech transcripts and user generated meta-data for which we compute BOW representations as well. It is reasonable to expect that semantic content of a video is strongly related to its genre, and if this semantic information was available it would make genre recognition simpler and more reliable. To this end, we used annotations for 345 semantic classes from TRECVID 2011 semantic indexing task to train semantic class detectors. Responses of these detectors were then used as features for genre recognition. The paper explains the approach in detail, it shows relative performance of the individual features and their combinations measured on MediaEval 2011 genre recognition dataset, and it sketches possible future research. The results show that, although, meta-data is more informative compared to the content-based features, results are improved by adding content-based information to the meta-data. Despite the fact that the semantic detectors were trained on completely different dataset, using them as feature extractors on the target dataset provides better result than the original low-level audio and video features.

1 INTRODUCTION

Genre is one of the most basic information which is available for movies, TV shows and other professional video content. It is presented to viewers in television programs, cinemas, rental shops and their on-line variants. It is important information a potential viewer takes into consideration when choosing the content to view.

Although the genre information is available for commercial content, it is not always supplied for user videos and other amateur content, or this information is not detailed or reliable enough. For example, Youtube allows users to choose one of 15 mutually exclusive categories without any suggestion based on the video content. Video sharing sites would benefit from automatic genre recognition when this information is not supplied by the user, and uploading videos could be made easier by suggesting genre to the user based on the content.

The approach presented in this paper was developed for MediaEval 2011 Genre Tagging Task (Larson et al., 2011; Hradis et al., 2011). The data provided for this task include Creative Commons videos downloaded from blip.tv, automatic speech recognition (ASR) transcripts (Gauvain et al., 2002), user supplied metadata for the videos and relevant social data from Twitter. Each video was assigned one of 26 exclusive genre categories. Some of the results reported in this paper are the runs submitted for evaluation while other results were acquired later on released test data using the official MediaEval evaluation methodology.

We approached the genre recognition task as a classification problem in a way which is consistent with the current trends in extraction of semantic information from video (Snoek et al., 2010; Smeaton et al., 2009). We relied on Support Vector Machine (SVM) to learn models of the individual genres based on several types of Bag of Word (BOW) representations. These classifiers, each using different type of features, were fused by Logistic regression to obtain final

decision rules. The information sources used in the experiments were still images extracted from video, audio, ASR transcripts and from user-supplied metadata. Local image features (van de Sande et al., 2010; Mikolajczyk and Schmid, 2005) were extracted from the images. The audio was transformed into spectrograms and from these local image features were extracted as well. The local features were transformed to BOW by codebook transform (van Gemert et al., 2010). Similarly, BOW representations were constructed from the ASR transcripts and metadata.

The approach outlined in the previous paragraph aims to map low-level visual and acoustic information directly to the video genre categories. Although some genres can be characterized by general visual and audio properties, such as dark colors, fast movement or dissonant and unsettling music (Brezeale and Cook, 2008), it is quite a leap from the the low level features to the very high level genre information. Having semantic information (e.g. objects, activities, environment type, screams and gun shots) extracted from the video would definitely aid in recognizing the genres. However, todays automatic methods for extraction of semantic information from video reach only moderate accuracy and require considerable resources and effort, especially, when many semantic classes are to be detected (Snoek et al., 2010; Smeaton et al., 2009). Even with the shortcomings of the automatic methods in mind, it is still reasonable to expect that they can provide some useful information for genre recognition. We explored this idea by training detectors of 345 semantic categories on TRECVID 2011 Semantic Indexing (SIN) task data, and by using results of these detectors as features for genre recognition.

Older approaches to video genre recognition are well summarized in a survey by Brezeale and Cook (Brezeale and Cook, 2008). Recently, You et al. (You et al., 2010) proposed a method for semantic video genre classification based on Gaussian Mixture Models, Hidden Markov Models and naive Bayes classifiers.

The paper is structured as follows. Section 2 describes our approach including feature extraction, classification and fusion, and it explains how the TRECVID semantic classifiers are used. The MediaEval data are described in Section 3. Experiments and their results are presented in Section 4. Finally, the paper is concluded in Section 5.

2 METHOD

As mentioned earlier, our approach relies on SVM classifiers, logistic regression for classifier fusion, and several BOW representations based on different modalities. Several types of features are combined by classifier fusion in order to get more robust and reliable classifications. The following text describes all the parts of the classification framework (see Figure1), and it introduces the semantic classifiers and explains how they are used for genre recognition.

2.1 Feature Extraction

For the purpose of extraction of visual features, 100 frames were extracted from each video. The frames were extracted equidistantly from the whole video. We did not use shot change information or any type of key-frame detection. The reason for this is that we intended to sample the videos as representatively as possible, and shot information or key-frame detection could favor certain type of content (e.g. fast dynamic scenes).

From each image, six types of local features were extracted. These six types were created by combining three image sampling methods and two descriptors. The first sampling method uses the Harris-Laplace scale invariant detector (Mikolajczyk, 2004) which localizes corner-like structures in scale-space. Rotation invariance was not used as it generally degrades results in image recognition problems. Further, two dense sampling schemes were employed. In both cases, circular regions were sampled on a regular grid with step of 8 pixels. The two schemes differ in the radius of the circular regions which was 8 pixels and 16 pixels.

The local image patches were parameterized using the original SIFT descriptor by Lowe (Lowe, 1999) and RGB-SIFT (van de Sande et al., 2010). The SIFT descriptor computes Histograms of Oriented Gradients (HOG) on a 4×4 grid centered on an image patch. The computed descriptors are vectors of 128 values created by concatenating the 16 histograms. The magnitude of a single pixel is distributed between neighboring histograms according to a spatial Gaussian filter which alleviates the boundary effect. The SIFT descriptor is invariant to shifts and scaling of the image intensity channel. It encodes the shape of an image patch while being resistant to small displacements and geometric transformations. The RGB-SIFT descriptor (van de Sande et al., 2010) computes SIFT independently on R, G and B image channels. For computational reasons, Principal Component Analysis was used to reduce dimensionality of the computed descriptors to 198.

To create feature vectors suitable for classification, codebook transform was used to translate the

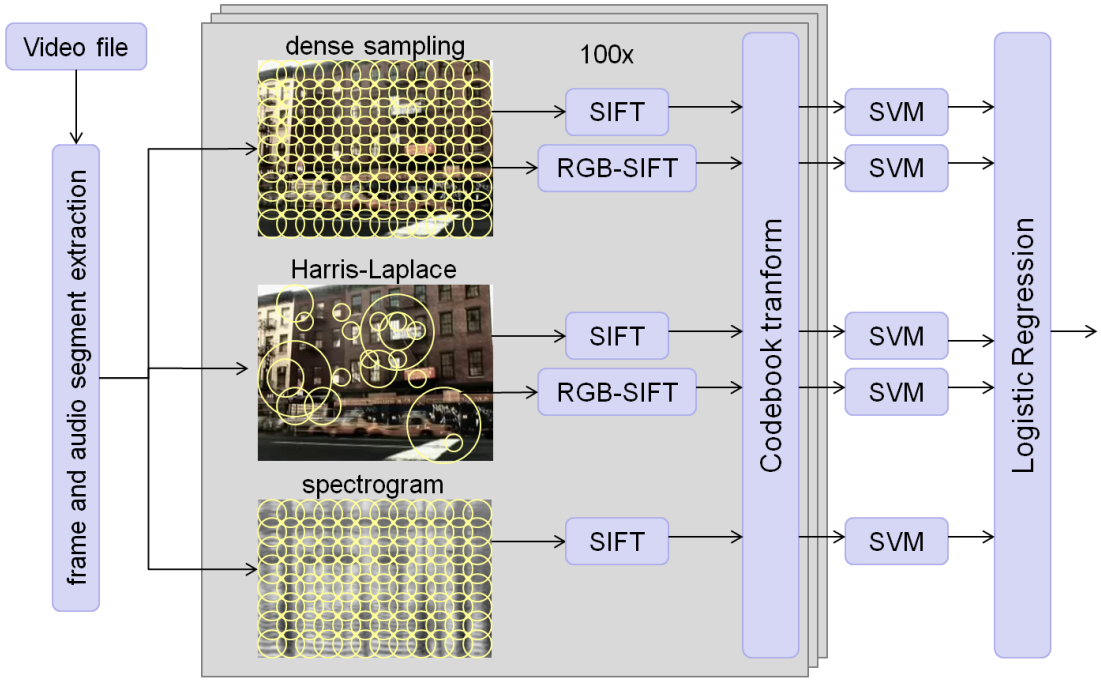

Figure 1: The processing pipeline: 100 frames and audio segments (spectrograms) are extracted from a video file. Spatial sampling is performed. Local patches are described by SIFT and RGB-SIFT. BOW representations are computed by codebook transform. SVM classifiers are trained for each representation and the classifiers are fused by logistic regression.

sets of local descriptors to BOW representations. Generally, codebook transform assigns objects to a set of prototypes and it computes occurrence frequency histograms of the prototypes. The prototypes are commonly called codewords and a set of prototypes is called a codebook. In our case, the codebooks were created by k-means algorithm with Euclidean distance. The size of the codebooks was 4096.

When assigning local features to codewords by hard mapping, quantization errors occur and some information is lost. This is especially significant in high-dimensional spaces, as is the case of the local patch descriptors, where the distances to several nearest codewords tend to be very similar. In the context of image classification, this issue was discussed for example by Gambert et al. (van Gemert et al., 2010) who propose to distribute local patches to close codewords according to codeword uncertainty. Computation of BOW with codeword uncertainty is defined for each codeword w from a codebook B as

$$UNC(w) = \sum_{p \in P} \frac{K(w,p)}{\sum_{v \in B} K(v,p)}, \qquad (1)$$

where P is a set of local image features and K is a kernel function. We use Gaussian kernel

$$K(w,w') = exp\left(-\frac{\|w-w'\|_2^2}{2\sigma^2}\right), \qquad (2)$$

where σ defines the size of the kernel. In our experiments σ was set to the average distance between two closest neighboring codewords from the codebook.

For parameterization of the audio information, an approach similar to parameterization of the visual information was used. The audio track was regularly segmented into 100 possibly overlapping segments. The length of the segments was 10 seconds and overlap was allowed as necessary. Mel-frequency spectrograms with 128 frequency bands, maximum frequency 8 KHz, window length 100ms and overlap 80ms were computed from these segments. Dynamic range of the spectrograms was reduced to fit 8-bit resolution. The spectrograms were then processed as images by dense sampling and SIFT descriptor. BOW representation was constructed for the spectrograms by codebook transform the same way as for images.

For classification, BOW histograms of the individual images and audio segments were averaged to get single BOW vector of each representation for each video.

From the metadata and the ASR data, XML tags were removed together with any non-alphabetical characters and words where lower-case character was followed by upper-case character were split. Stemming was not performed on the data. Although, the data includes several Dutch, French and Spanish videos, we did not employ any machine translation, as

the ratio of the non-English videos is relatively small and it should not seriously influence results. For each video, separate word occurrence counts for metadata and ASR were collected.

All feature vectors were normalized to unit length for classification.

2.2 Classification Scheme

Although, the data in MediaEval genre tagging task is multi-class (a video is assigned to a single class), the evaluation metric is Mean Average Precision, and the genre recognition problem is in general multi label - one video may belong to several genres, e.g. Sci-Fi and comedy. As a result, we build classifiers for each genre separately and independently.

The classification structure has two levels. The first level consists of linear SVM classifiers each based on a single BOW representation. These classifiers are then fused by logistic regression to produce robust estimates of the genres.

SVM (Cortes and Vapnik, 1995) is often used for various tasks in image and video classification (Le et al., 2011; van de Sande et al., 2010; van Gemert et al., 2010; Snoek et al., 2010; Smeaton et al., 2009). SVM has four main advantages. It generalizes well, it can use kernels, it is easy to work with, and good-quality SVM solvers are available. Although non-linear kernel have been shown to perform better in image recognition (Perronnin et al., 2010), we selected linear kernel due to the very small training set size. Radial Basis Function kernels which are usually used (Perronnin et al., 2010; van de Sande et al., 2010; van Gemert et al., 2010; Snoek et al., 2010) introduce an additional hyper-parameter which has to be estimated in cross-validation on the training set. Estimating this parameter together with the SVM regularization parameter could prove to be unreliable on the small dataset.

The single SVM regularization parameter was estimated by grid search with 5-fold cross-validation if enough samples for particular class were available. The objective function in the grid search was Mean Average Precision and the same parameter was used for all genre classes for a particular BOW representation.

Due to the fact that no validation set was available, we had to re-use the training set for Logistic Regression fusion. To keep the classifiers from overfitting, we trained the Logistic Regression on responses of the 5 classifiers learned in cross-validation with the estimated best value of the SVM hyper-parameter. Each classifier computed responses on the part of the data which it was not trained on. This way, no knowledge of a particular video was used to compute response on that video. Before fusion, classifier responses were normalized to have zero mean and unit standard deviation. Multinomial L2-regularized logistic regression was used for the fusion. The regularization parameter was estimated by the same grid search and cross-validation procedure as in the case of the linear SVMs. Considering the different nature of the available features, the video and audio classifiers (see Section 2.1) were fused separately and the classifiers using semantic features (see further in Section 2.3) were fused separately as well. Finally, the two classifiers created by fusion were fused again with the classifiers based on the ASR transcripts and metadata. In this second fusion, single set of weights was computed for the different modalities and these were used for all genres in order to limit overfitting.

2.3 Semantic Detectors in Genre Recognition

The TRECVID[1] 2011 SIN task provided a training dataset consisting of approximately 11,200 videos with total length of 400 hours. The duration of the videos ranges from 10s to 3.5 minutes. The source of the videos is Internet Archive[2]. The videos were partitioned into 266473 shots (Ayache et al., 2006) which are represented by a corresponding keyframe. The 500 semantic classes proposed by TRECVID organizers were annotated by active learning[3] (Ayache and Quénot, 2007; Ayache and Quénot, 2008). Total 4.1M hand-annotations were collected and this produced 18M annotations after propagation using relations (e.g. Cat implies Animal). For 345 classes, the annotations contained more than 4 positive instances. Examples of the classes are Actor, Airplane Flying, Bicycling, Canoe, Doorway, Ground Vehicles, Stadium, Tennis, Armed Person, Door Opening, George Bush, Military Buildings, Researcher, Synthetic Images, Underwater and Violent Action.

Using the TRECVID SIN task data, 345 semantic classifiers were trained. These classifiers use the same eight BOW feature types and the same SVM classifiers as described in Section 2.1 and Section 2.2. Further details on these classifiers can be found in (Beran et al., 2011) together with the results achieved in TRECVID 2011 evaluation.

We applied these 345 classifiers to the extracted images and audio segments and created feature representations for the videos by computing histograms

[1] http://trecvid.nist.gov/

[2] http://www.archive.org/

[3] http://mrim.imag.fr/tvca/

Genre	#videos
art	4
autos and vehicles	1
business	7
citizen journalism	11
comedy	6
conferences and other events	2
default category	40
documentary	3
educational	19
food and drink	4
gaming	4
health	7
literature	6
movies and television	6
music and entertainment	2
personal or auto-biographical	5
politics	45
religion	16
school and education	2
sports	7
technology	27
the environment	1
the mainstream media	5
travel	4
videoblogging	10
web development and sites	3

Table 1: Genre categories and their distribution in training set.

of their responses. The histograms consisted of 8 equidistant bins with the outer bins set to 5% quantiles. The dimension of the resulting feature vectors obtained by concatenating the histograms of individual semantic classes was 2760.

3 DATA

The data released for MediaEval 2011 Genre Tagging Task (Larson et al., 2011) consist of Creative Commons videos downloaded from blip.tv. The data is split into separate training and testing sets. The training set contains 247 videos and the test set contains 1728 videos. The genre categories and the distribution of classes in the training set is shown in Table 1.

The metadata includes information about the videos supplied by uploaders of the video including among others title, description, uploader login name and tags.

MAP of a random system on the test set should be 0.046.

Features		TV11
DENSE16_CSIFT	0.193	0.194
DENSE16_SIFT	0.149	0.178
DENSE8_CSIFT	0.178	0.201
DENSE8_SIFT	0.138	0.187
HARLAP_CSIFT	0.186	0.178
HARLAP_SIFT	0.170	0.174
SPECTRUM_DENSE16_SIFT	0.183	0.167
SPECTRUM_DENSE16_SIFT	0.175	0.189
COMBINED	0.254	0.276

Table 2: Mean average precision achieved by individual types of features. Line COMBINED contains results of fusion of all the presented features.

4 EXPERIMENTS AND RESULTS

In experiments, we focused mostly on the performance the individual content-based features, and on how much fusion of the metadata with the content-based features improves over the results achieved by using metadata alone. The metadata is in general very informative as, for example, one uploader usually uploads videos of only very small number of genres.

The results of visual and audio features are shown in Table 2. In the table, dense sampling is denoted as DENSE16 for patch radius of 16 pixels and DENSE8 for radius of 8 pixels (see Section 2.1. HARLAP stands for the Harris-Laplace detector. The descriptors are denoted as SIFT and CSIFT for the RGB-SIFT. The audio features are indicated by SPECTRUM. TV11 represents the classifiers based on responses of TRECVID semantic detectors.

The results show that the semantic classifiers on average provide better results. The same is true for fusion of the the audio and video features where classifiers using directly the low-level features achieve MAP 0.254 and those using semantic features achieve 0.276.

Table 3 compares results of metadata, ASR transcripts, fusion of content-based features, and fusion of all features including metadata. Metadata by itself gives MAP 0.405. On the other hand results of ASR are much lower. With MAP 0.165 it is slightly worse than the individual audio and video features. By combining all content-based features MAP 0.304 was achieved. This is still significantly lower than for the metadata. However, it is important to realize that the content-based features do not include any information generated by humans. By combining metadata with the content-based features, the results are improved by 0.046 reaching MAP 0.451.

Four runs were submitted to MediaEval 2011 using the presented approach with several differences (Hradis et al., 2011). The most notable differ-

Figure 2: Frames randomly selected from MediaEval 2011 genre tagging task data. The source videos are released under Creative Commons license.

Features	MAP
All including metadata	0.451
Metadata	0.405
All Content-based	0.304
ASR	0.165

Table 3: Mean average precision achieved by fusion of all features, metadata alone, all content based features (audio, video and ASR), and ASR alone.

Run	MAP
RUN1	0.165
RUN3	0.346
RUN4	0.322
RUN5	0.360

Table 4: Mean average precision on test set achieved by the runs submitted to MediaEval 2011.

ence was that weights for the classifier fusion were set by hand. When fusing the audio and video features, uniform weights were used. **RUN1** used only ASR. **RUN3** combined all features with the weight of ASR and METADATA increased to 2.5. **RUN4** combined the low-level audio and video features, ASR and metadata. Here the weights of ASR and metadata were set to 1.25. **RUN5** combined semantic features, ASR and metadata with the same weights as in RUN4. The results of these runs are show in Table 4.

The best purely content-based method submitted to MediaEval 2011 achieved MAP 0.121 (Ionescum et al., 2011). Very successful were methods focusing on metadata and information retrieval methods. The best result was MAP 0.56 (Rouvier and Linares, 2011). This result was reached by explicitly using IDs of the uploaders and the fact that uploaders tend to upload similar videos. Other than that, the approach classified the data by SVM on metadata, ASR transcripts and audio and video features.

5 CONCLUSIONS

The presented genre recognition approach achieves good result on the datasets used in the experiments. The results could be even considered surprisingly good considering the small size of the training set used. However, it is not certain how the results would generalize to larger and more diverse datasets.

Although the metadata is definitely the most important source of information for genre recognition, the audio and video content features improve results when combined with the metadata. Compared to the metadata, content-based features achieve worse results, but they do not require any human effort.

The semantic features for classification improve over the low-level features individually, as well as, when combined.

ACKNOWLEDGEMENTS

This work has been supported by the EU FP7 project TA2: Together Anywhere, Together Anytime ICT-2007-214793, grant no 214793, and by BUT FIT grant No. FIT-11-S-2.

REFERENCES

Ayache, S. and Quénot, G. (2007). Evaluation of active learning strategies for video indexing. *Signal Processing: Image Communication*, 22(7-8):692–704.

Ayache, S. and Quénot, G. (2008). Video Corpus Annotation Using Active Learning. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R., editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 187–198. Springer Berlin / Heidelberg.

Ayache, S., Quénot, G., and Gensel, J. (2006). CLIPS-LSR Experiments at TRECVID 2006. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID.

Beran, V., Hradis, M., Otrusina, L., and Reznicek, I. (2011). Brno University of Technology at TRECVid 2011. In *TRECVID 2011: Participant Notebook Papers and Slides*, Gaithersburg, MD, US. National Institute of Standards and Technology.

Brezeale, D. and Cook, D. J. (2008). Automatic Video Classification: A Survey of the Literature. *IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews*, 38(3):416–430.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Gauvain, J.-L., Lamel, L., and Adda, G. (2002). The LIMSI Broadcast News transcription system. *Speech Communication*, 37(1-2):89–108.

Hradis, M., Reznicek, I., and Behun, K. (2011). Brno University of Technology at MediaEval 2011 Genre Tagging Task. In *Working Notes Proceedings of the MediaEval 2011 Workshop*, Pisa, Italy.

Ionescum, B., Seyerlehner, K., Vertan, C., and Lamber, P. (2011). Audio-Visual Content Description for Video Genre Classication in the Context of Social Media. In *MediaEval 2011 Workshop*, Pisa, Italy.

Larson, M., Eskevich, M., Ordelman, R., Kofler, C., Schmeideke, S., and Jones, G. J. F. (2011). Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Workshop*, Pisa, Italy.

Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *Learning*, pages 1–4.

Lowe, D. G. (1999). Object Recognition from Local Scale-Invariant Features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA. IEEE Computer Society.

Mikolajczyk, K. (2004). Scale & Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86.

Mikolajczyk, K. and Schmid, C. (2005). A Performance Evaluation of Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630.

Perronnin, F., Senchez, J., and Xerox, Y. L. (2010). Large-scale image categorization with explicit data embedding. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2297–2304, San Francisco, CA.

Rouvier, M. and Linares, G. (2011). LIA @ MediaEval 2011 : Compact Representation of Heterogeneous Descriptors for Video Genre Classication. In *MediaEval 2011 Workshop*, Pisa, Italy.

Smeaton, A. F., Over, P., and Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Divakaran, A., editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin.

Snoek, C. G. M., van de Sande, K. E. A., de Rooij, O., Huurnink, B., Gavves, E., Odijk, D., de Rijke, M., Gevers, T., Worring, M., Koelma, D. C., and Smeulders, A. W. M. (2010). The MediaMill TRECVID 2010 Semantic Video Search Engine. In *TRECVID 2010: Participant Notebook Papers and Slides*.

van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2010). Evaluating Color Descriptors for Object and Scene Recognition. {*IEEE*} *Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596.

van Gemert, J. C., Veenman, C. J., Smeulders, A. W. M., and Geusebroek, J. M. (2010). Visual Word Ambiguity. *PAMI*, 32(7):1271–1283.

You, J., Liu, G., and Perkis, A. (2010). A semantic framework for video genre classification and event analysis. *Signal Processing Image Communication*, 25(4):287–302.