Audio-Visual Classification Video Browser

David Scott[1], Zhenxing Zhang[1], Rami Albtal[1], Kevin McGuinness[1], Esra Acar[2], Frank Hopfgartner[2], Cathal Gurrin[1], Noel E. O'Connor[1] and Alan F. Smeaton[1]

[1] Dublin City University, Glasnevin, Dublin 9, Ireland.
[2] Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany

Abstract. This is our third participation in the Video Browser Showdown, having participated in 2012 with a handheld system and 2013 with a desktop system, the latter placing second overall and first with respect to novice users. Building on the experience that we gained while participating in this event, we compete in the 2014 showdown with a more advanced browsing system. This paper provides a short overview of the features and functionality of our new system.

Video Segmentation: We segment the video into shots based on techniques outlined by Pickering et al.[6]. Further, we extract multiple key frames for each shot to represent the shot in the graphical user interface. A segmentation into shots allows us to provide a quick overview over different scenes within a video, a strategy which proved successful in the TRECVid known item search task. It is likely that some key frames will be very similar to other key frames of the same shot. To avoid including these similar frames, we remove duplicates by comparing the global color layout of all key frames within each shot.

Visual Concept Classification: We use trained models to classify the visual content based on concepts such as person, landscape or buildings. We use the judgements from the classification to act as ranked list when used alone. We also use the concept lists to either filter or boost content when used in tandem with other searches.

The initial classifiers were trained on TrecVid 2013 Semantic Indexing task (SIN) training set [9]. In this task 60 concepts were requested to be identified in the shots of the SIN video dataset. Three visual content descriptors were used: two types of Opponent SIFT Bag of Visual words, the first is based on dense sampling, the second on Haaris-Laplace sampling [10]; the third descriptor is a concatenation of a normalized RGB Histogram and a normalized Gabor transform. A Support Vector Machine (RBF-euclidean distance kernel)[1] was trained for each of the three descriptor and each of the 60 concepts. To provide a judgement about the existence of a particular concept in a shot, the correspondent three classifiers of that concept were used to evaluate the visual features extracted from the shot, then a weighted sum of the three judgement scores generated by the classifiers is performed to provide a final score. An initial framework for feature extraction and classification parameter evaluation and

optimisation is developed and tested in this task; the framework is designed to be extendible to work on a large scale data, it is installed on the machines of the Irish Centre for High-End Computing (ICHEC)[3].

Audio Concept Classification: We generate models to detect audio concepts such as explosions, gunshots and screams. In a manner similar to the visual concept classification, we use audio concept lists to either filter or boost content when used in tandem with other searches. The classifiers are trained on the MediaEval 2013 Violent Scenes Detection task (VSD) training set [2] using high-level audio concept annotations provided in the VSD dataset. We employ Mel-Frequency Cepstral Coefficients (MFCC) features as low-level audio features. For the representation of video shots, we use a Bag-of-Audio Words (BoAW) approach based on MFCC with a sparse coding scheme. We adopt the dictionary learning technique presented in [5]. In the coding phase, we construct the sparse representations of audio signals by using the LARS algorithm [3]. In order to generate the final sparse representation of video shots which are a set of MFCC feature vectors, we apply the *max-pooling* technique. A Support Vector Machine (SVM) with an RBF kernel is trained for each audio concept using sparse audio representations. In order to provide a judgment about the existence of a particular concept in a shot, the probability estimates of SVM models are used. Normally, in a basic SVM, only class labels or scores are output. The class label results from thresholding the score, which is not a probability measure. The scores output by the SVM are converted into probability estimates using the method explained in [12].

Visual Similarity Search: The aim of visual similarity search is to offer our system the ability to find the most visual relevant shots to a given shot query and then provide the most possible shots for users to identify. Supposing the searching topic is happened in front of a special scene, we could easy filter out the shots which did not happened in that scene.

Different from the approach of our previous participation [7] using locally aggregation descriptors (VLAD) and nearest neighbours searching, we followed the route of [8] and built a linear discriminative object classifier for each query image. The main benefit of this approach is that a unique weighting score will be learnt to determine the most discriminative visual features for retrieval from training data which contains one positive data and many negative data.

For each keyframe, local feature descriptors are extracted and an aggregation descriptor is generated to represent it. A large size of negative training set is created and reused for every classifier training. In the online process, the same dimensional descriptor is extracted to the search query image and a linear classifier is trained by using the open source library [4]. Finally, each video shot from dataset can be sorted by calculating the inner product of weighting vector from classifier and their feature vectors.

[3] `http://www.ichec.ie/`

Face Browsing and Search: We anticipate that providing functionality to allow users to get a high-level overview of all the human faces appearing in a video will be useful for queries involving people. To this end, we provide a face view that shows all the faces found in the video, and allows users to quickly navigate to the locations in the video in which selected faces appear. We use the Viola-Jones face detector [11] to first locate faces in the videos, and then to cluster these faces by using agglomerative techniques. This clustering also allows face-based search to be easily implemented: when the user chooses to search for similar faces on a given key frame, all images associated with the clusters containing any faces that appear in the key frame can be retrieved and displayed.

References

1. Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.
2. Claire-Helene Demarty, Cedric Penet, Markus Schedl, Bogdan Ionescu, Vu Lam Quang, and Yu-Gang Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.
3. Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
4. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
5. Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.
6. Marcus J. Pickering and Stefan M. Rüger. Evaluation of key frame-based retrieval techniques for video. *Computer Vision and Image Understanding*, 92(2-3):217–235, 2003.
7. David Scott, Jinlin Guo, Cathal Gurrin, Frank Hopfgartner, Kevin McGuinness, Noel E. O'Connor, Alan F. Smeaton, Yang Yang, and Zhenxing Zhang. Dcu at mmm 2013 video browser showdown. In *MMM (2)*, volume 7733 of *Lecture Notes in Computer Science*, pages 541–543. Springer, 2013.
8. Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics (TOG) (Proceedings of ACM SIGGRAPH ASIA)*, 30(6), 2011.
9. Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.
10. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
11. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. pages 511–518, 2001.
12. Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.