The Multimedia Understanding Group at TRECVID-2010

Christos Diou, George Stephanopoulos and Anastasios Delopoulos

Multimedia Understanding Group
Dept. of Electrical and Computer Engineering
Aristotle University of Thessaloniki, Greece
Email: {diou,stephan}@mug.ee.auth.gr,adelo@eng.auth.gr

Abstract

This is a report of the Multimedia Understanding Group participation in TRECVID-2010, where we submitted full runs for the Semantic Indexing (SIN) task. Our submission aims at experimentally evaluating three research items, that are important for work that is currently in progress. First, we examine the use of bag-of-words audio features for video concept detection, with noisy and/or low-quality video data. Although audio is important for some concepts and has shown promising results at other datasets, the results indicate that it can also lead to a decrease in performance when the quality is low and the negative examples are not adequately represented. We also explore the possibility of using a cross-domain concept fusion approach for reducing the number of dimensions at the final classifier. The corresponding experiments show, however, that when drastically reducing the number of dimensions the effectiveness drops. Finally, we also examined a transformation of the feature space, using a set of functions that are parametrically constructed from the data.

Semantic Indexing Runs

1. F_A_MUG-AUTH_1: Early fusion of all available low-level features.
2. F_A_MUG-AUTH_2: Early fusion of all available low-level features except audio.
3. F_A_MUG-AUTH_3: Feature space transformation based on a sample of the training data.
4. F_C_MUG-AUTH_4: Concept fusion using a limited number of base classifiers.

1 Introduction

This year's video collection is different from previous TRECVID [1] datasets in many ways: The collection is larger and very diverse in both the type of thematic content and the audiovisual quality, while the number of concepts is also larger, compared to the "High-level feature extraction" task [2] of previous years. The runs that we submitted aim at exploring

1. How audio features behave on a collection with such diverse characteristics.

2. If a cross-domain concept fusion approach with a small number of classifiers can reach the same effectiveness as the original features.

3. To evaluate the behavior of a feature space transformation method that is still in development.

In the following, we provide a short presentation of the methods used, for low level feature extraction, fusion and the classifier. We then detail the experimental setup of the runs and discuss the results.

2 Low-level features

For low-level feature extraction, our system uses the visual, audio and text modality. The visual features have been detailed in previous submissions [3, 4] and include

- A feature based on the MPEG-7 [5] Color Structure Descriptor [6] (CSD, 64-d).
- A feature based on the dominant color of the image (DCOLOR, 30-d). For the extraction of the feature octrees are used for color reduction [7] along with a set of reference images (as in [8]) that are compared with the input keyframe using the Earth Mover's Distance [9].
- A feature based on a histogram of the Hough transform angles (HOUGH, 24-d).
- A feature based on the description of the edges of keyframe regions using the Integrated Weibull distribution (WBL, 120-d) and the comparison of distributions with a set of reference images, as in [8].
- The color SIFT feature described in [10], using dense sampling of keypoints and the HSV colorspace (LOCAL, 4000-d).

For the text feature (TEXT), one vocabulary of the most frequent words was created for each high-level concept, based on the shots that were positive in the development set. The text feature is simply a bag-of-words histogram of these words in each shot.

For audio (AUDIO, 1000-d), we extracted Mel-Frequency Cepstral Coefficients (MFCC) from the audio channel and created 1000 clusters using the k-means algorithm. These clusters were treated as "audio words" and for each shot, a histogram of audio words is constructed using the euclidean distance of shot's MFCCs from the cluster centers. A "soft weighting" scheme is employed for the weights, that uses the fuzzy c-means weighting function,

$$v_i = \frac{1}{\sum_j \left(\frac{d(\mathbf{c}_i,\mathbf{m})}{d(\mathbf{c}_j,\mathbf{m})}\right)^{2/(u-1)}} \tag{1}$$

where $\mathbf{c}_i$ are the cluster centers and $\mathbf{m}$ the input point in the MFCC space, while $u > 1$ is a parameter that controls the "steepness" of the weighting function. If $u \to 1$ then this function provides a hard assignment of points to cluster centers. In the experiments, a set of 1000 audio words was used and the assignment was carried out using Equation (1) and $u = 1.5$.

3 Fusion methods

The simplest fusion approach is early fusion, where the original feature vectors are concatenated, possibly after removal of outliers, normalization and weighting. A more complex fusion scheme that is applied is cross-domain concept fusion: A feature vector is constructed by the outputs of a set of base classifiers that have been trained at a foreign domain. The classifiers are constructed for each feature separately or even using subsets of features. This concept fusion method can therefore provide multiple levels of detail for the description of the target domain. In [11] we study this approach and two criteria are introduced for the selection of the base classifiers that are most informative for the target domain.

In this year's experiments, we examine if selecting a relatively small number of base classifiers can achieve comparable effectiveness with early fusion, while reducing the dimensionality at the final classifier stage.

Another direction that we explore is the application of a feature space transformation, where the base classifiers of the concept fusion approach are replaced by a set of "basis functions" that are defined parametrically by the data. Each point $\mathbf{f}$ of the original feature space is transformed to a vector $\mathbf{b}$ that contains the values of these functions,

$$\mathbf{b} = \begin{bmatrix} g_1(\mathbf{f}) & \dots & g_N(\mathbf{f}) \end{bmatrix} \tag{2}$$

The aim of the transformation is to simplify the binary classification problem that can lead to improved effectiveness/generalization.

4 Semantic indexing runs

We used the training data that was created by the collaborative annotation effort organized by LIG and LIF [12, 13]. In order to control the computational complexity of the training stage we did not use the entire set for training, but applied a sampling strategy: For each concept, all N_p positive examples are kept and random sampling is applied for selecting $N_n = \max(10^4, 2N_p)$ negative examples. The sampling stage is applied once and all runs use the same samples so that results are comparable.

The following runs were submitted:

Run 1: Early fusion of all features (LOCAL, AUDIO, CSD, DCOLOR, HOUGH, WBL, TEXT).

Run 2: Early fusion of all features, except audio (LOCAL, CSD, DCOLOR, HOUGH, WBL, TEXT).

Run 3: Transformation of the fused feature of Run 2. One transformation is created for each concept, using the sampled training set to construct the functions g_i of Equation (2).

Run 4: Cross-domain concept fusion using a small number of base classifiers. Using the TRECVID-2005/LSCOM ([14, 15]) training data we construct multiple classifiers for all features and select a common set of 700 base classifiers for all concepts. These base classifier outputs are used for training.

Runs 1 and 2 compare the use of the AUDIO feature in this year's collection. With Run 4, we examine if it is possible to use a common set of features (the classifier outputs) with fewer dimensions and still achieve results comparable to Run 1. Finally, Run 3 is a first evaluation of the feature space transformation method that we are developing. The results for the concepts that were evaluated by the TRECVID organizers are provided at Table 1.

Comparing Runs 1 and 2, it is easy to see that for most concepts, the audio feature did not help. This can be attributed to the following reasons:

1. The high number of dimensions of the audio feature led to a decrease in performance, without always compensating this with discriminative power for concept detection.

2. Audio cues may be misleading for some concepts. For example, it is hard to see how audio would help the detection of "Hand". However, given the limited number of examples, audio information may be associated with "Hand" by the classifier, leading to false detection results.

3. The sampling strategy we applied reduced the number of negative examples and it is possible that some errors in the class boundaries were introduced, that became apparent with the audio feature.

Table 1: Results for the set of concepts that were evaluated, for all runs.

Concept	TV10 ID	Run 1	Run 2	Run 3	Run 4
Airplane_Flying	4	0.0230	**0.0240**	0.0100	0.0050
Animal	6	**0.0120**	0.0100	0.0090	0.0030
Asian_People	7	0.0040	**0.0090**	0	0
Bicycling	13	0.0010	**0.0100**	0.0070	0
Boat_Ship	15	0.0320	**0.0460**	0.0130	0.0400
Bus	19	**0.0030**	**0.0030**	0	**0.0030**
Car_Racing	22	0.0050	**0.0290**	0.0070	0.0040
Cheering	27	**0.0180**	0.0150	0.0050	0.0170
Cityscape	28	0.0640	**0.0770**	0.0230	0.0450
Classroom	29	0.0020	0.0020	0	0.0010
Dancing	38	0.0130	**0.0210**	0.0010	0.0080
Dark-skinned_People	39	0.0370	0.0160	**0.0540**	0.0130
Demonstration_Or_Protest	41	0.0210	**0.0360**	0.0110	0.0130
Doorway	44	0.0460	**0.0480**	0.0320	0.0230
Explosion_Fire	49	0.0230	0.0230	**0.0320**	0.0040
Female-Human-Face-Closeup	52	**0.0490**	0.0370	0.0430	0.0480
Flowers	53	0.0140	**0.0280**	0.0150	0.0040
Ground_Vehicles	58	0.0660	0.0900	0.0780	**0.1140**
Hand	59	0.0120	**0.0210**	0.0170	0.0190
Mountain	81	0.1140	**0.1840**	0.1190	0.0830
Nighttime	84	0.0510	**0.0530**	0.0290	0.0110
Old_People	86	**0.0210**	0.0180	0.0120	**0.0210**
Running	100	0.0110	0.0250	0.0080	**0.0390**
Singing	105	**0.0520**	0.0270	0.0130	0.0120
Sitting_Down	107	0.0050	**0.0060**	0.0020	0.0050
Swimming	115	0.2430	**0.3250**	0.0560	0.0040
Telephones	117	0.0040	**0.0070**	0	0
Throwing	120	0.0020	0.0010	0.0030	**0.0220**
Vehicle	126	0.0550	0.0840	**0.0860**	0.0710
Walking	127	0.0330	0.0600	0.0310	**0.0720**
Mean infAP		0.0345	**0.0445**	0.0239	0.0235

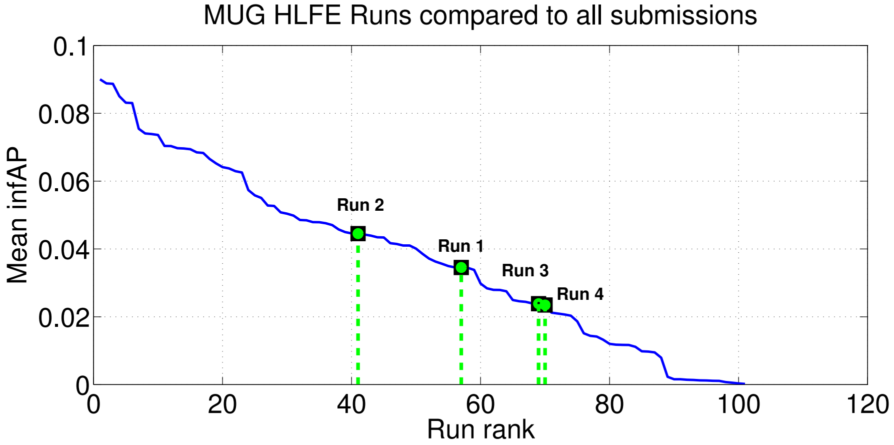

Figure 1: Our runs compared to all TRECVID-2010 submissions.

There were cases, however, where audio information was helpful, such as the concept "Singing". Overall, this experiment shows that (i) not all concepts benefit from audio so a careful selection of the concepts where audio should be used is advised, (ii) audio features seem to work adequately with fewer number of dimensions (e.g., in [4] we used 500) and (iii) when using audio features, a large number of negative examples may be needed so that audio information is not falsely associated with the concept.

Runs 1 and 4 show that for this set, the selection of the 700 base classifiers (for a feature space of more than 5238 dimensions, not including the TEXT feature) leads to a decrease in effectiveness. This decrease is on average 0.011 in infAP, according to Table 1. One will therefore need to more classifiers and increase the detail of the description for the target feature space in order to reach or exceed the results of Run 1 with concept fusion. Finally, Run 3 shows that the feature transformation approach can lead to results similar to concept fusion, without the need for base classifiers.

In Figures 1 and 2 we show how this year's submissions compare to the systems of other participants.

References

[1] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, (New York, NY, USA), pp. 321–330, ACM Press, 2006.

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

[3] C. Diou, C. Papachristou, P. Panagiotopoulos, G. Stephanopoulos, N. Dimitriou, A. Delopoulos, H. Rode, R. Aly, A. P. de Vries, and T. Tsikrika, "VITALAS at TRECVID-2008," in *Proceedings of the TRECVID 2008 Workshop*, 2008.

[4] C. Diou, G. Stephanopoulos, N. Dimitriou, P. Panagiotopoulos, C. Papachristou, A. Delopoulos, H. Rode, T. Tsikrika, A. P. de Vries, D. Schneider, J. Schwenninger, M.-L. Viaud, A. Saulnier, P. Altendorf, B. Schröter, M. Elser, A. Rego, A. Rodriguez, C. Martínez, I. n.

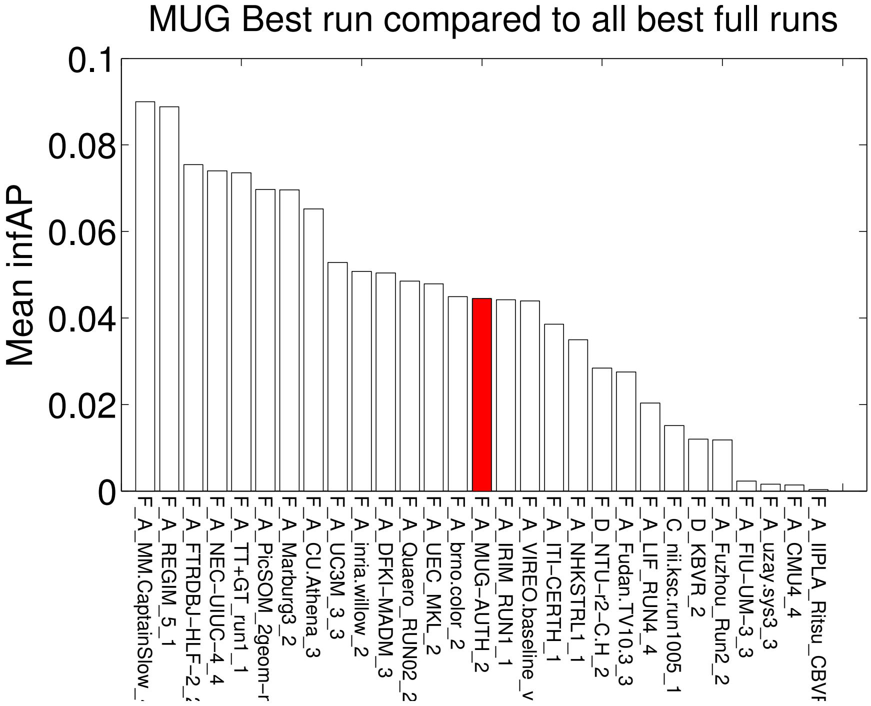

Figure 2: Our best run compared to the best run of each participant that submitted full runs.

Etxaniz, G. Dupont, B. Grilhères, N. Martin, N. Boujemaa, A. Joly, R. Enficiaud, A. Verroust, S. Selmi, and M. Khadhraoui, "VITALAS at TRECVID-2009," in *Proceedings of the TRECVID 2009 Workshop*, 2009.

[5] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface.* Wiley, 2002.

[6] D. Messing, P. van Beek, and J. Errico, "The mpeg-7 colour structure descriptor: image description using colour and local spatial information," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1, pp. 670–673, 2001.

[7] M. Gervautz and W. Purgathofer, "A simple method for color quantization: Octree quantization," in *New Trends in Computer Graphics*, Springer Verlag, Berlin, 1988.

[8] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, "Robust scene categorization by learning image statistics in context," in *Proceedings of the International Workshop on Semantic Learning Applications in Multimedia, CVPRW*, 2006.

[9] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, November 2000.

[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*, vol. 32, pp. 1582–1596, September 2010.

[11] C. Diou, G. Stephanopoulos, P. Panagiotopoulos, C. Papachristou, N. Dimitriou, and A. Delopoulos, "Large-scale concept detection in multimedia data using small training sets and

cross-domain concept fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, to appear.

[12] G. Quénot, A. Tseng, B. Safadi, and S. Ayache, "Trecvid-2010 collaborative annotation." `http://mrim.imag.fr/tvca2010`, 2010.

[13] A. S. and G. Quénot, "Video corpus annotation using active learning," in *Proceedings of the 30th European Conference on IR Research*, pp. 187–198, 2008.

[14] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton, "TRECVID 2005 - An overview," in *Proceedings of the TRECVID 2005 Workshop*, 2005.

[15] L. Kennedy and A. Hauptmann, "LSCOM lexicon definitions and annotations version 1.0." DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report, 217-2006-3, 2006.