Eurecom at TRECVID 2009
High-Level Feature Extraction

Feng Wang and Bernard Merialdo

Multimedia Communications Dept.
Institute Eurecom
Sophia-Antipolis, France
{*Feng.Wang, Bernard.Merialdo*}*@eurecom.fr*

1 Abstract

This year Eurecom submitted 5 runs for the High-Level Feature Extraction task. Below are brief descriptions of these submissions.

- **A_Eurecom_fuse_base**: this run fuses different visual features, including Bag-of-Visual-Words (BoW), Color Moment (CM), Wavelet Texture (WT), Edge Histogram (EH) and Local Binary Pattern (LBP).
- **A_Eurecom_rerank1**: this run reranks the results of the previous run **Eurecom_fuse_base** based on the prior probability of the concept in shots and videos.
- **A_Eurecom_rerank2**: this run updates the concept score of the run **Eurecom_fuse_base** based on context knowledge.
- **A_Eurecom_specific**: this run updates the concept score of the run **Eurecom_rerank2** with the output of several specific detectors for face, person and bicycle respectively.
- **A_Eurecom_visual_audio**: this run fuses the run **Eurecom_specific** with two audio features, i.e. MFCC and Audio Spectral.

With these runs, first we try to evaluate the performance of different visual and audio descriptors for high-level feature extraction; second, we investigate the use of context and video knowledge in reranking the detection results. The evaluation results show that the reranking schemes significantly improve the performance of the concept detectors, especially for those context-dependent concepts. Figure 1 illustrates the framework of our system and how the five runs are generated. In the following sections, we describe the details of descriptor extraction, and the approaches for classification and reranking.

We also participated in the joint IRIM submission. Our contribution is covered in the IRIM notebook paper.

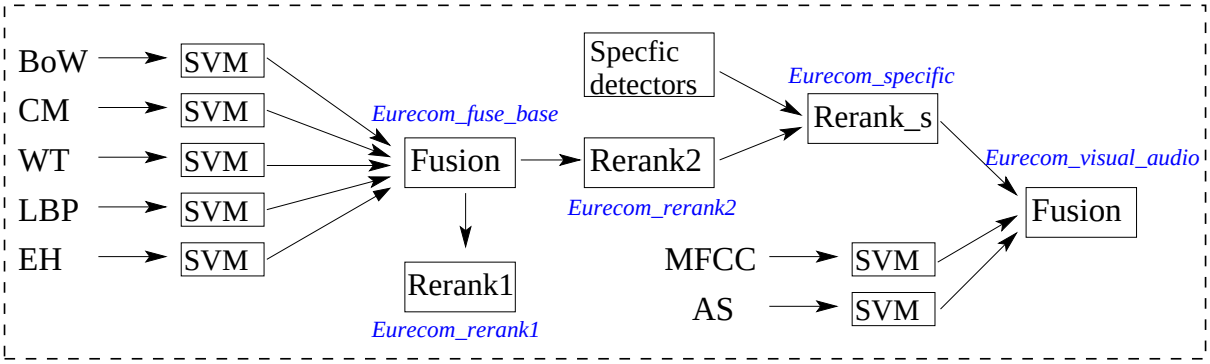

Figure 1: Overview of our system for HLF task. **BoW:** Bag-of-Visual-Words; **CM:** Color Moment; **WT:** Wavelet Texture; **LBP:** Local Binary Pattern; **EH:** Edge Histogram; **MFCC:** Mel-Frequency Cepstral Coefficients; **AS:** Audio Spectral. **Rerank_1:** reranking with video knowledge; **Rerank_2:** reranking with context knowledge; **Rerank_s:** reranking with specific detectors.

2 Data and Descriptors

In all the 5 runs, we use the SV07 data and the collaborative annotations of the 20 concepts from LIG for development. The keyframes are provided by IRIM [11]. The following descriptors are extracted from the keyframes (or shots for audio descriptors):

Bag-of-Visual-Word (BoW): In each keyframe, local interest points (LIPs) are detected using Difference of Gaussian (DoG) and Laplacian of Gaussian (LoG) [3], and then described by SIFT (Scale-Invariant-Feature-Transformation) [2]. All the extracted LIPs in development data are clustered into 500 groups by employing k-means clustering to form a visual vocabulary with 500 words. Given a keyframe, each LIP is then assigned to the nearest visual word and statistics are collected over the frame to build a feature vector of 500-bin histogram. In this work, we used the softwares for LIP detection and description implemented by VIREO group [8].

Color Moment (CM): For each keyframe, the first 3 moments of 3 channels in *Lab* color space over 5×5 grids are calculated, and aggregated into a 225-d feature vector.

Wavelet Texture (WT): A given keyframe is split into 3×3 grids and each grid is represented by the variances in 9 Haar wavelet sub-band to form a $81 - d$ feature vector.

Edge Histogram (EH): We extract the MPEG-7 edge histogram descriptor which represents the spatial distribution of five types of edges, namely four directional edges (one horizontal, one vertical, and two diagonal edges) and one non-directional edge for 16 local regions in each keyframe, and form a 80-bin feature vector.

Local Binary Pattern (LBP): Local binary pattern describes the local texture information around each point [5], which has been proven effective in object recognition. We employ the implementation in [7] to extract and combine the LBP features with three different radius (1, 2, and 3) and get a 54-bin feature vector.

Mel-Frequency Cepstral Coefficients (MFCC): We download the MFCC feature from IRIM [11] which is contributed by IRIT. The feature is calculated as the average and variance

of MFCC coefficients.

Audio Spectral (AS): We download the Audio Spectral features from IRIM [11] which is contributed by GIPSA.

3 Baseline

Our baseline run is produced by fusing the visual descriptors listed in Sec 2. We use a Binary Support Vector Machine (SVM) [6] for classification. For each visual descriptor, a SVM model is trained based on the development data for the prediction of each concepts. The parameters of SVM are learned by cross-validation between SV07 development and test data.

The concept scores from different descriptors are then fused with weighted linear fusion. The weights of each feature for each concept are learned based on the development data. The result of descriptor fusion is included in the run **Eurecom_fuse_base**.

4 Reranking

4.1 Reranking with Video Knowledge

The presence of some concepts is dependent on the nature of the video. For instance, a shot from a music video may contain the concepts *Person-playing-a-music-instrument*, *Singing* or *People-dancing* with higher probability. With the concept scores predicted by SVM, we are able to estimate the category of the given video and thus rerank the shots according to the video knowledge.

Given a shot s in video v, the best ranking of shots for concept c is

$$P(s|c) = \frac{P(c|s) \cdot P(s)}{P(c)} \tag{1}$$

where $P(c|s)$ is the concept score calculated by the SVM and fusion, and $P(c)$ is constant for all shots given the concept c and can be ignored. With knowledge of shots and videos,

$$P(s) = \frac{P(s|v) \cdot P(v|c)}{P(c)} \tag{2}$$

Thus, we have

$$\begin{aligned} P(s|c) &\propto P(c|s) \cdot P(s|v) \cdot P(v|c) \\ &= P(c|s) \cdot P(s|v) \cdot \frac{\sum_{s' \in v} P(c|s')}{\sum_{\tilde{s}} P(c|\tilde{s})} \\ &\propto P(c|s) \cdot \frac{1}{N_v} \sum_{s' \in v} P(c|s') \\ &= P(c|s) \cdot P_{v,c} \end{aligned} \tag{3}$$

where N_v is the shot number in v, and $P_{v,c}$ is the average concept scores over all shots in v.

In Equation 3, we can see the shot ranking score is proportional to $P_{v,c}$ which, to some extent, indicates the category of the video v. Since the presence of some concepts (*e.g.* Chair and Doorway) are not closely related to the video nature, we just apply Equation 3 to the concepts for which the average precision can be significantly improved during development phase. The result after reranking the baseline with Equation 3 is included in the run **Eurecom_rerank1**.

4.2 Reranking with Context Knowledge

Equation 3 assumes the knowledge of video and shot is available. In the reranking, we use the predicted concept scores in the baseline as the known knowledge. However, this knowledge is dependent on the performance of the SVMs. Intuitively, the highest concept scores are usually more reliable, and thus we just employ the information from those shots and videos with the highest concept scores.

We calculate the average and variance of video concept confidence as

$$\begin{aligned} \mu_{v,c} &= \frac{1}{V} \cdot \sum_{v} P_{v,c} \\ \sigma_{v,c} &= \sqrt{\frac{\sum_{v}(P_{v,c} - \mu_{v,c})^2}{V}} \end{aligned} \tag{4}$$

where V is the total video number in the dataset. We just consider video v for reranking that satisfies the following condition

$$P_{v,c} - \mu_{v,c} > \lambda_c \cdot \sigma_{v,c} \tag{5}$$

where λ_c is a significance factor estimated for different concepts using development data. In the selected videos, we then find the shots with highest concept scores

$$P(c|s) - \mu_{v,c} > \lambda_s \cdot \sigma_{v,c} \tag{6}$$

where λ_s is a significance factor for shots. A shot s selected by Equations 5 and 6 contains the concept c with high confidence. For some concepts such as *Person-playing-a-musical-instrument* and *Singing*, the neighboring shots of s also contain the same concept with high probability. Thus, we update the concept scores of its neighboring shot s' by

$$P(c|s') = P(c|s') + \beta \cdot \frac{1}{d(s,s')} \cdot P(c|s) \tag{7}$$

where $d(s,s')$ is the distance from s' to s, and β is an influence factor estimated from development phase. This approach is used to rerank the baseline and the result is included in the run **Eurecom_rerank2**.

5 Reranking with Specific Detectors

We download three specific detectors, i.e. face detector[9], person detector and bicycle detectors [10], and run them on the extracted keyframes. These detectors are used to rerank and improve

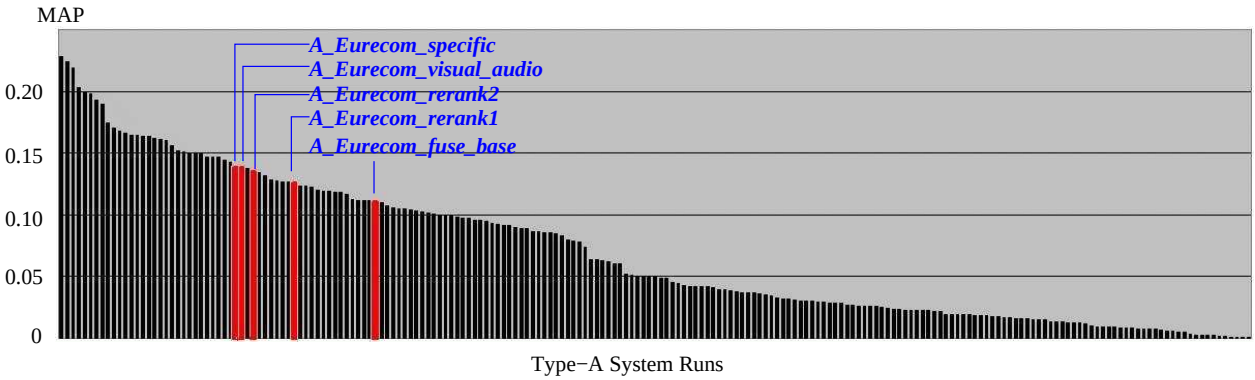

Figure 2: Performance of our submitted type-A runs and all official type-A HLF system runs.

the concept detection results. Given a shot s, a specific detector outputs a confidence score f_s, which indicate the probability that the shot contains a face, a person or a bicycle. We update the ranking score $P(s|c)$ as

$$P'(c|s) = P(c|s) * (1 + \alpha * f_s) \tag{8}$$

where α is an influence factor which can be estimated for each specific detector on each concept using development data. Equation 8 is applied to those concepts for which the AP can be significantly improved by using the specific detectors in development phase. This approach is applied to the results of the run **Eurecom_rerank2** to produce the run **Eurecom_specific**.

6 Results and Discussions

Figure 2 shows the evaluations results for our submitted runs and all the type-A runs, and Figure 3 shows the performance of our submitted run for different concepts. In the baseline (**Eurecom_fuse_base**), different visual descriptors are fused. According to our experiments on evaluating single descriptors, the local descriptor BoW gets the highest MAP, which has proven effective in other works [1, 4]. Meanwhile, by combining other features, significant improvement can still be achieved. Different descriptors are good at capturing different information in the images, *e.g.* **CM** for color, and **EH** for edge distribution.

Compared with the baseline, reranking with video knowledge (**Eurecom_rerank1**) achieves an improvement of 13.5%. This improvement is mainly contributed by those concepts that are dependent on the video categories including *People_dancing* (202% improvement), *Person-playing-a-musical-instrument* (126%), *Singing* (53.6%), and *Boat_Ship* (34.0%).

By reranking the results with context knowledge as described in Sec 4.2, the MAP is improved by 21.9%. Similarly, the improvement is also contributed by some concepts including *People_dancing* (373% improvement), *Person-playing-a-musical-instrument* (101%), *Singing* (72.5%), *Classroom* (64.5%), and *Boat_Ship* (34.0%). As discussed in Sec 4.2, by using only the highest confidence scores, this reranking scheme performs better than reranking with video knowledge. An disadvantage is that more parameters have to be set to achieve good results.

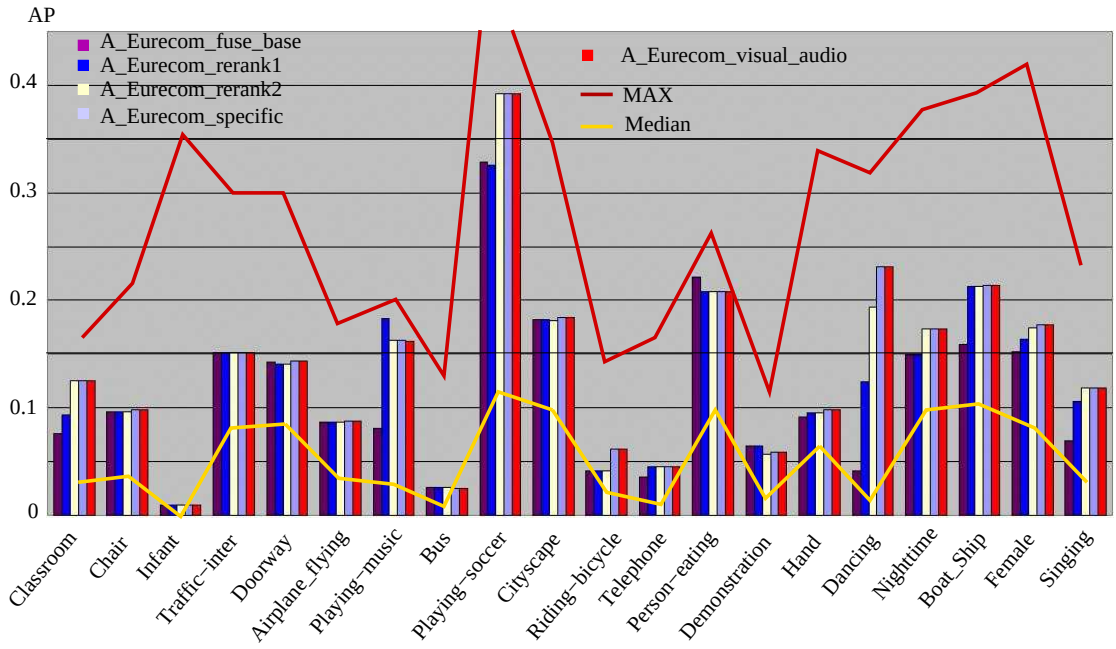

Figure 3: Average precisions of our submitted runs for different concepts. The two lines show the max and median performance for all type-A system runs.

By applying specific detectors **Eurecom_specific**, the MAP is slightly improved by 2.8%. Bicycle detector improves the AP of concept *Person-riding-a-bicycle* by 51.2% from 0.041 to 0.062. Person detector improves the AP of concept *People-dancing* by 19.1% from 0.194 to 0.231. Face detector is employed to rerank the concept *Female-human-face-closeup*. However, the result is only slightly improved by 1.7%. This may be because the result for this concept has been good enough, where most of the detected images already contain faces.

In the last run, we fuse the visual and audio descriptors for two concepts *Person-playing-a-musical-instrument* and *Singing*. However, due to a programming problem, the AP is slightly decreased. In our development phase, the APs of the two concepts were improved by around 5-10% by combining audio descriptors.

7 Conclusion

We have presented our approach for high-level feature extraction task. In this first attempt on this task, we evaluated the performance of different visual and audio descriptors. Furthermore, we tried several ranking schemes to employ video context knowledge and also specific detectors, which are demonstrated to be useful for improving the detection results.

8 Acknowledgements

EURECOM's research is partially supported by its industrial members: BMW Group, Cisco, Monaco Telecom, Orange, SAP, SFR, Sharp, STEricsson, Swisscom, Symantec, Thales.

References

[1] S. F. Chang, J. He, Y. G. Jiang, E. Khoury, C. W. Ngo, A. Yanagawa, and E. Zavesky, "Columbia University/VIREO-CityU/IRIT: TRECVID2008 High-Level Feature Extraction and Interactive Video Search", *TRECVID Workshop*, 2008.

[2] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *Int. Journal of Computer Vision*, vol. 60, no. 2, 2004.

[3] K. Mikoljczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. Journal of Computer Vision*, vol. 60, pp. 63–86, 2004.

[4] C. W. Ngo, Y. G. Jiang, X. Wei, W. Zhao, F. Wang, X. Wu, H. Tan, "Beyond Semantic Search: What You Observe May Not Be What You Think", *TRECVID Workshop*, 2008.

[5] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971-987.

[6] LIBSVM. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[7] Local binary pattern. http://www.ee.oulu.fi/mvg/page/home.

[8] VIREO group. http://vireo.cs.cityu.edu.hk.

[9] W. Kienzle, G. Bakir, M. Franz and B. Scholkopf, "Face Detection - Efficient and Rank Deficient" *Advances in Neural Information Processing Systems* vol. 17, pg. 673-680, 2005

[10] Inria object detection. http://www.irisa.fr/vista/Equipe/People/Ivan.Laptev.html.

[11] IRIM website. http://mrim.imag.fr/irim/wiki/doku.php.