EURECOM and ECNU at TrecVid 2010: The Semantic Indexing Task

Miriam Redi, Bernard Merialdo
Multimedia Department, EURECOM
Sophia Antipolis, France
{redi,merialdo}@eurecom.fr

Feng Wang
East China Normal University
Shanghai, China
fwang@cs.ecnu.edu.cn

October 22, 2010

1 Abstract

This year EURECOM and ECNU participated together at the TRECVID Semantic Indexing Task. We built four different systems for the light (10 concepts) submission. Three of our runs are functionally similar to the system used by EURECOM for last year's High Level Feature Extraction task (see [6] for further details).

We keep as a basic run (Fusebase) the best-performing system from 2009, testing how such system performs on the new dataset; we then improve the EURECOM_Fusebase by adding a global descriptor, originally built for scene recognition, and proved to be effective in the TRECVID context for spatially-independent concepts like "Nighttime". We then experiment with a multi-modal analysis, combining the visual features with the textual metadata that have been provided with the 2010 video database. As last run, we try a new system based on Hamming Embedding and Weighted Visual words. The runs are composed as follows:

1. **EURECOM_Fusebase** This run fuses a pool of visual features, namely the Sift descriptor, the Color Moments global descriptor, the Wavelet Feature and the Edge Histogram. On top of this, a face detector and a re-ranking method based on the video knowledge are applied, according to the 4^{th} run that EURECOM presented in the 2009 Edition.

2. **EURECOM_Gist** This run increases the set of visual features with a global descriptor presented by Torralba et al. in [5] called the *gist* descriptor.

3. **EURECOM_Metadata** This run adds to the previous run the information mined from the textual metadata files related to each video.

4. **EURECOM_Weight_HE** this run combines Hamming Embedding and Weighted Bag-of-Visual-Words

Beside this participation, EURECOM took part in the joint IRIM submission; systems details are included in the IRIM notebook paper.

The remainder of this paper briefly describes the content of each run (Sec 2-5), including feature extraction, fusion and reranking methods. In Section 6 results are commented and discussed.

2 EURECOM Basic Run: EURECOM_Fusebase

This run is composed by three main modules: first, a model is built by combining a pool of visual features; a face detector is then merged into the system; finally, a re-ranking method based on context knowledge is applied.

1. ***Visual Feature Extraction and Fusion***: in this stage, 4 different features are computed. For each feature a Support Vector Machine is trained to predict the presence of a concept c in a keyframe s of the test set. The choice of the descriptors is based on their effectiveness on the training set. For each keyframe the following descriptors are extracted:

   - **Sift with Bag of Words** Two sets of interest points are identified using different salient points detectors:
     - Difference of Gaussian
     - Hessian-Laplacian Detector

     Each of these key points is then described with the SIFT [4] descriptor, using the VIREO system [1]. A K-means algorithm clusters a subset of the training set descriptors in a vocabulary of n visual words. Then, for each SIFT point in a keyframe, the nearest neighbor in the vocabulary is calculated; based on this statistics a n-dimension feature vector is built collecting the number of points in the image that can be approximated by the n^{th} visual word. Experiments on the training set suggested to choose $n = 500$.
   - **Color Moments** This global descriptor computes, for each color channel in the LAB space, the first, second and third moment statistics.
   - **Wavelet Feature** This texture-based descriptor calculates the variance in the Haar wavelet sub-bands for each window resulting from a 3×3 division of a given keyframe.
   - **Edge Histogram** The MPEG-7 edge histogram describes the edges' spatial distribution for 16 sub-regions in the image.

   The features extracted are then used as input to feed a Support Vector Machine (SVM) (see implementation details in [2]) that creates, for each concept c, a model based on each feature extracted from the training data. Such model is then used to detect the presence of c in a new sample s based on each feature. The classifiers parameters are selected via exhaustive grid search: their value is chosen based on the Mean Average Precision

maximization on the development set. We have now, for each concept c and keyframe s, 5 feature-specific outputs that we call $p_n(c,s)$, $n = 1,\ldots,5$. We fuse such scores with weighted linear fusion in order to obtain a single output, namely $p_v(c,s)$, that represents the probability of the concept c in the keyframe s given the set of visual features.

2. ***Face Detector***: $p_v(c,s)$ (see Figure 1) is then interpolated with the output, defined as $p_f(c,s)$, of a face specific detector ran on the s. The interpolation is computed as follows: $p(s|c) = p_v(c,s)(1 + \alpha \cdot p_f(c,s))$, where the parameter α is estimated by maximizing the Mean Average Precision of the development set.

3. ***Re-Ranking based on video and context knowledge***: as we did last year, $p(s|c)$ is updated based on a quantity P_v, namely the average concept score over all video shots, according to the equation $p(c|s) = p(s|c) + P_v$. Furthermore, the concept scores of the neighboring shots of s are updated based on $p(c|s)$ (futher details can be found in [6]).

We performed step 2 and 3 only for the concepts for which adding such modules was introducing a significant improvement in the final MAP (in the development stage).

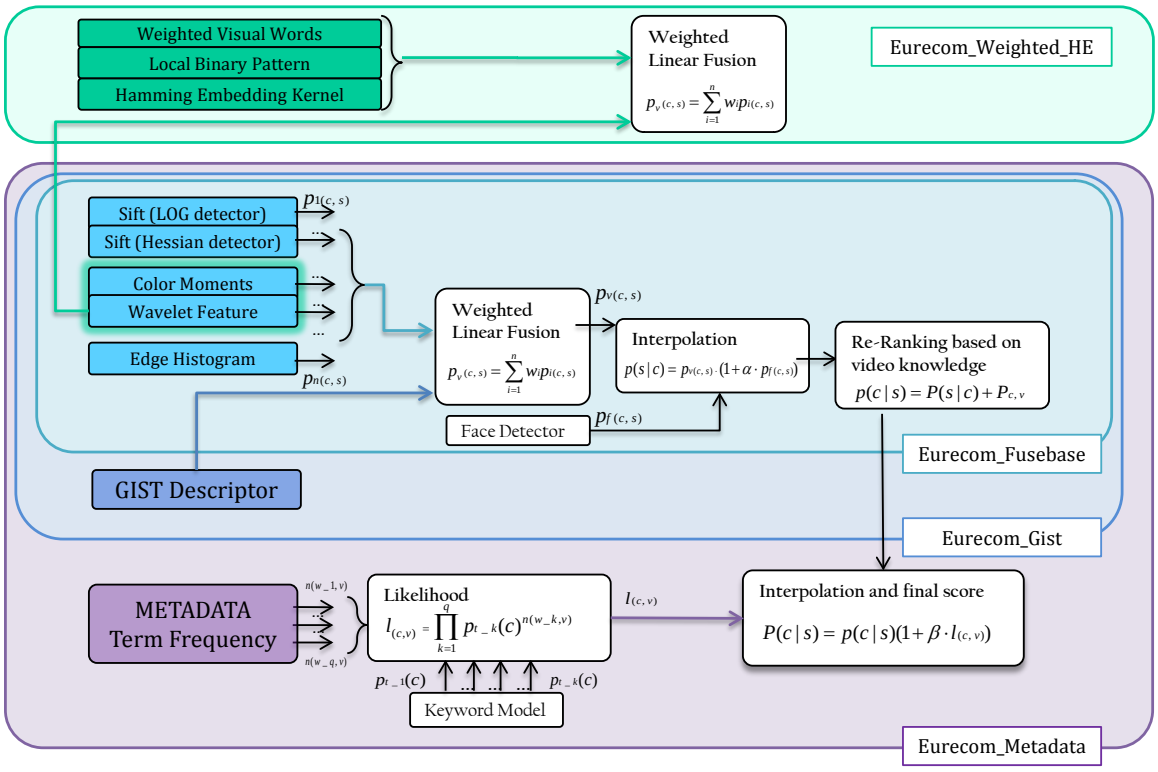

Figure 1: Framework of our system for the semantic indexing task

3 EURECOM Second Run: EURECOM_Gist

In this run, the Visual Feature Extraction and Fusion module of the basic run is improved by adding a new global descriptor in the feature pool, namely the *gist* descriptor, proposed by Torralba et al. in [5]. Originally built as a feature for natural scene recognition, we chose *gist* for its good performances in global concept detection (e.g. Cityscape, Nighttime). Such feature exploits the properties of the Fourier Spectrum to represent the spatial envelope of the image/keyframe, i.e. the general structure of the scene. As for the other visual features, the extracted descriptor for each keyframe is used as input for an SVM, and the corresponding output is fused with the other visual feature-based concept scores via linear fusion.

4 EURECOM Third Run: EURECOM_Metadata

Here a textual feature module is added to the previous visual-only run.

This year, a set of XML-formatted metadata is available with each video. We use the Term Frequency statistics to create a model for these textual descriptions: on the training set, for each concept c we compute the quantities $p_{t_k}(c)$, i.e. the probability for word t_k to appear in correspondence with concept c. We compute such statistics in a reduced set of fields of the XML metadata file, chosen based on their effectiveness in the global system performances, namely "title", "description", "subject" and "keywords'.

Given this model, on a new test video v we compute the cardinality $n(w, v)$, where w is a term that appears in the metadata file corresponding to v. We then compute the likelihood $l(c, v)$, between the test video textual feature and each concept-based text model. Such values are then used to update the output of the EURECOM_Gist run, obtaining, for each shot $s \in v$,

$$P(c|s) = p(c|s)(1 + \beta \cdot l(c, v))$$

The value β is estimated on the development data.

5 EURECOM Fourth Run: EURECOM_Weight_HE

This run fuses the Color Moments and Wavelet features computed for run 1-3 with Sift features. The key modules here rely on the techniques used to process the raw SIFT features: weighted visual words and Hamming embedding.

5.1 Weighting Informativeness of BoW

In the current BoW based approaches, for different concepts, each visual word is treated equally. The discriminative ability of those informative visual words could be seriously reduced. In this run, we employ an approach to measure the importance of each visual word for given concepts. This is achieved by iteratively updating the weights to push the SVM kernel towards the optimal one. For this purpose, kernel alignment score (KAS) is used to measure the discriminative ability

of SVM kernel. The problem is then to maximize the KAS score by optimizing the weights of different visual words. Finally, the resulting weights are used in a modified kernel for concept detection. The details can be found in [7].

5.2 Hamming Embedding

During the construction of the visual vocabulary, SIFT descriptor space is quantized into Voronoi cells corresponding to different visual words [3]. Such kind of quantization simply treats all points falling in one cell as identical which results in an inaccurate distance measure between image samples.

In this run, we propose Hamming Embedding kernel to alleviate information loss problem of BoW features in video concept detection. Hamming Embedding (HE) was originally employed in [3] for similar image search and copy detection. For each keypoint quantized by visual vocabulary, a binary signature is further generated encoding its location information inside the cell. The distance between two keypoints in the same cell can be roughly estimated by the Hamming distance between their binary signatures. We revise HE as a kernel for discriminative classification.

6 Results Analysis

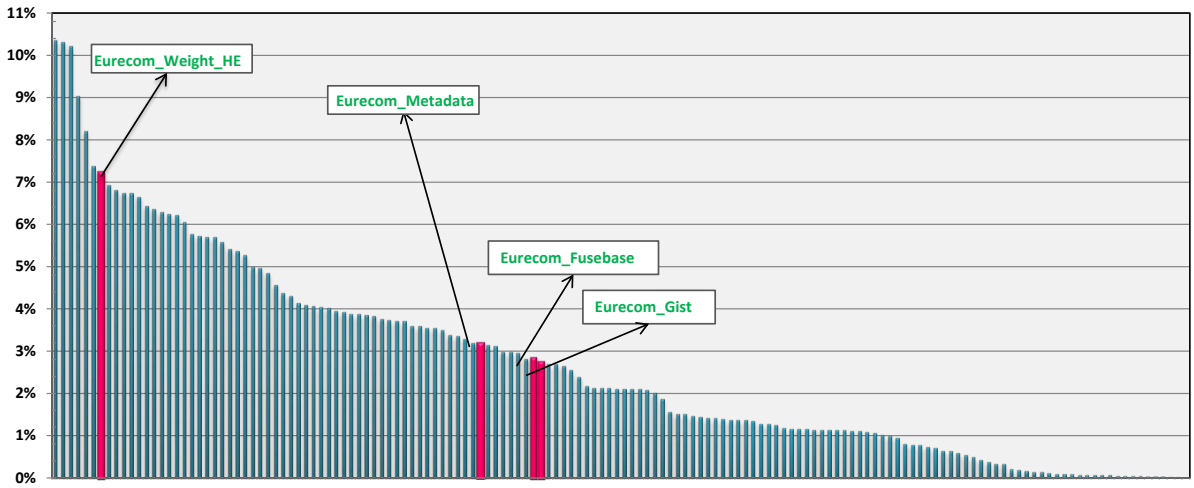

Figure 2: Evaluation results for our submitted runs

In Figure 2, the performances (MAP) of the various systems submitted for the light SIN task are presented.

The first three runs were based on classical visual features and textual information, as mentioned in Sec. 2-4. Shown in Fig 3 are the concept-specific final performances; for run 1,2,3 the expected MAP (i.e. the value expected from the training phase) is as well presented. Compared to the estimated Mean Average Precision , the performances on the test set decrease of about

60% for the basic run and 70% for the Eurecom_Metadata. Such systems have indeed been tuned on a specific subset of data, causing overtraining and performance decrease. n-fold cross validation and negative examples separation could have avoided this failure.

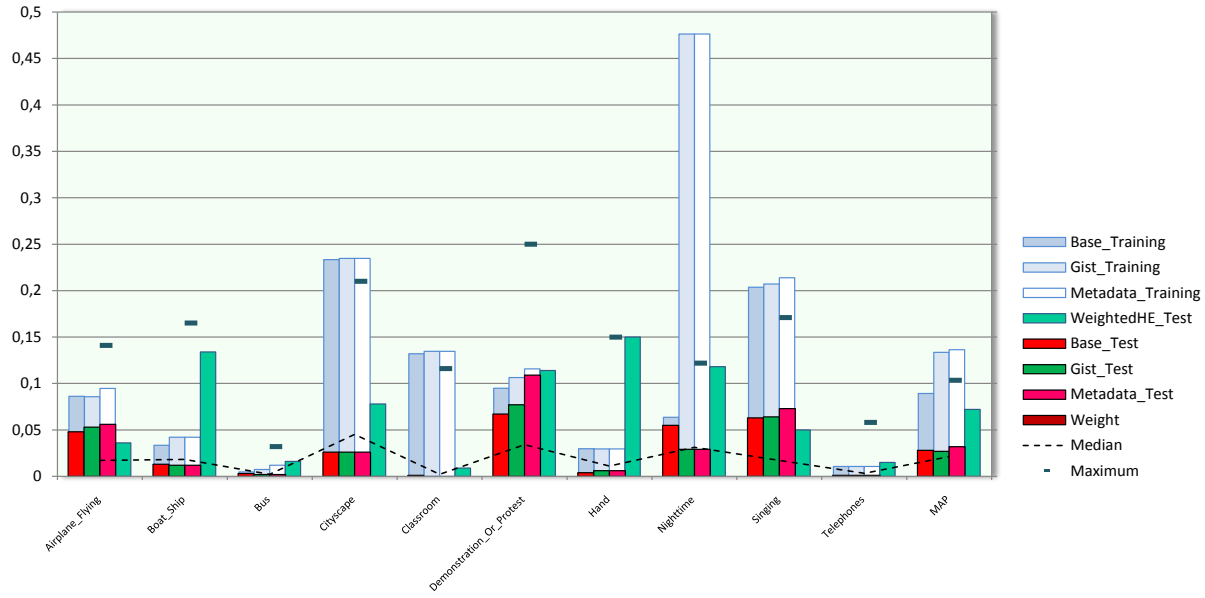

Figure 3: Framework of our system for the semantic indexing task, run 1-3

We can see in Fig 3 that, by adding the Gist descriptor in the pool of visual features, the system performs slightly worse; this is mainly due to the fact that, in test, the average precision for the concept *Nighttime* decreases of about 50%, while the predicted difference between Eurecom_Fusebase and Eurecom_Gist on the development (overtrained) data was around +600%. However, for other concepts, e.g. Demonstration_or_Protest (+ 15%) or Airplane_Flying (+ 10%), the *gist* descriptor contributes positively to the final MAP.

The use of textual data combined on top of the Eurecom_Gist run improves the MAP by 11,7%. Only 3 out of 10 concept scores (i.e. *Airplane_Flying* (+ 17%), *Demonstration_Or_Protest*(+ 62%), *Singing* (+ 16%), exactly the concepts for which the textual features have been extracted) contribute to the increase of the global performances.

The Eurecom_Weighted_HE is shown to be quite effective for local concepts (e.g. Boat_Ship and Hand). The reason for that is that such run relies on a different set of global features and, moreover, that it uses improved local features, increasing their discriminative power using the described techniques. Our development estimations showed that Weighting and HE brought about 11% and 15% improvement respectively.

7 Conclusions

This year EURECOM and ECNU presented together a set of systems for the Semantic Indexing Task. Three out of four runs aimed at improving a system similar to the Eurecom 2009 sub-

mission. We showed that by adding a global descriptor, namely the GIST descriptor, the final MAP increases for some concepts. We also found that adding textual features extracted from the metadata improves the visual-only based system. Hamming embedding and weighted visual words added to a set of global features are also proved to be very effective for local concepts detection.

References

[1] Vireo group in http://vireo.cs.cityu.edu.hk/links.html.

[2] C. Chang and C. Lin. LIBSVM: a library for support vector machines. 2001.

[3] M. D. H. Jegou and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European Conf. on Computer Vision*, 2008.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[6] F. Wang and B. Merialdo. Eurecom at TRECVID 2009 High-Level Feature Extraction.

[7] F. Wang and B. Merialdo. 'weighting informativeness of bag-of-visual-words by kernel optimization for video concept detection'. In *ACM Workshop on Very-Large-Scale Multimedia Corpus, Mining and Retrieval*, 2010.