EURECOM at TrecVid 2014: The Semantic Indexing Task

Usman Niaz, Bernard Merialdo, Claudiu Tanase
Multimedia Department, EURECOM
Sophia Antipolis, France
{Usman.Niaz,Bernard.Merialdo}@eurecom.fr

October 8, 2014

1 Abstract

This year EURECOM participated in the TRECVID 2014 Semantic INdexing (SIN) Task [12] for the submission of four different runs for 60 concepts. Our submission builds on the runs submitted last year at the 2013 SIN task, the details of which can be found in [9]. In 2014, two runs are trained using the TRECVID data only, these two runs only differ by the number of visual descriptors which are considered. The third run includes classifiers trained on the ImageNet corpus, the fourth run also applies our uploader model. All runs are trained on annotations provided by the IRIM collaborative effort [4].

When compared with last year system, our runs use a larger set of visual features, and larger visual dictionaries to provide a finer representation of the visual/clustering space and increase the precision of the retrieval system. Like in previous years submissions we add a global descriptor to visual features capturing salient details or gist of a keyframe. We further benefit from metadata information by including an uploader bias to increase the scores of videos uploaded by same users. As in 2013, we have used a new training algorithm based on a combination of PEGASOS [16] and Homogeneous Kernel Maps [19] which allows the simultaneous construction of several models, therefore reducing the time needed to train the system.

Our four runs are organized as follows:

1. **Run4**: This is our basic run which fuses a pool of 11 visual features, namely SIFT [7] descriptors extracted through log and hessian methods with 2K vocabularies, dense sampling, and dense color pyramid [18] with 10K vocabularies, a Saliency Moments feature [13], a Color Moments global descriptor, a Wavelet Feature, a Edge Histogram, a texture-based Local Binary Pattern feature, and a spatio-temporal edge histogram descriptor [17]. The order of the Homogeneous Kernel Maps is 5, so each scalar component is translated into a 11 dimension vector. The classifiers are linear SVM on Homogeneous Kernel maps trained

with the PEGASOS algorithm. A linear fusion combines the results of the classifiers to provide the final score.

2. **Run3**: This run uses the same descriptors as the previous one, but adds various sizes of vocabularies for the SIFT-based features, 500 and 1K for the point-based, 1K, 4K and 10K for the dense, with or without spatial pyramids, to reach a total of 28 descriptors. All these descriptors are trained on the TRECVID data only.

3. **Run2**: This run add two extra descriptors, composed of semantic vectors obtained by applying pretrained classifiers on the TRECVID videos. One set of classifiers is based on the pretrained CAFFE Neural Network [6], which is a Deep Neural Net, the other is based on linear SVM applied on Fisher Vectors [15]. Both sets contain 1000 classifiers that have been trained on the ImageNet data. Those classifiers are applied unchanged on the TRECVID video data, and the 1000 scores for each shot are assembled into a 1000 dimension semantic vector which is used as the feature vector for this shot.

4. **Run1**: This run takes the results of Run2 and applies our uploader model [8]. The uploader model uses statistics about the concepts that occur in the videos posted by each uploader. Those statistics are computed on the training data and used as prior probabilities on the test data. This process can be applied because a substantial number of videos in the test data has been uploaded by people who also uploaded videos in the training data.

Beside this participation, EURECOM took part in the collaborative IRIM and VideoSense submissions; the details of those systems are included in the respective papers.

The remainder of this paper briefly describes the content of each run (Sec 2-5), including feature extraction, classifier training and fusion methods. Figure 1 gives an overview of the relationship between the 4 runs. In Section 6 results are commented and discussed.

2 EURECOM Basic Run: Run4

This run comprises 11 visual features ranging from local features to global image descriptions. In this stage, the following features are computed.

- **Color Moments** This global descriptor computes, for each color channel in the LAB space, the first, second and third moment statistics on 25 non overlapping local windows per image.

- **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.

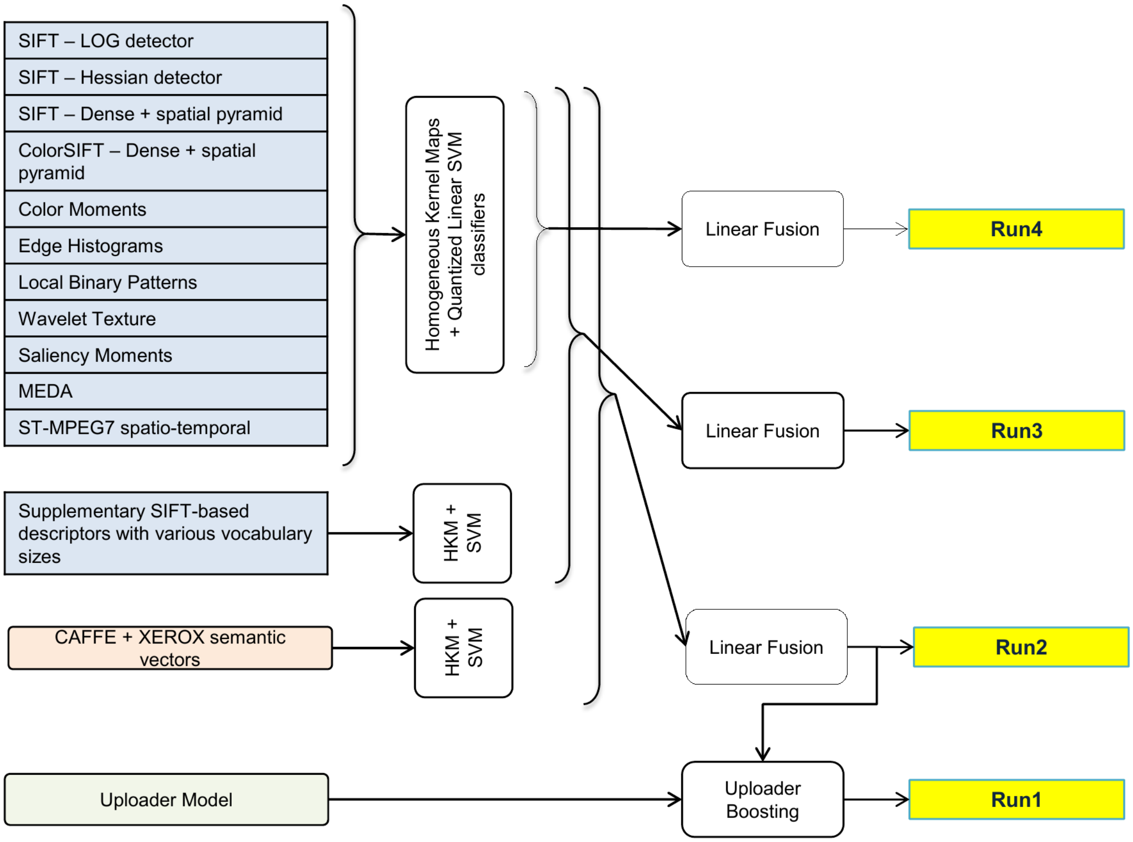

Figure 1: Framework of our system for the Semantic INdexing task

- **Local Binary Pattern (LBP)** Local binary pattern describes the local texture information around each point [10], which has been proven effective in object recognition. We employ the implementation in [2] to extract and combine the LBP features with three different radius (1, 2, and 3) and get a 54-bin feature vector.

- **SIFT from keypoints** Two sets of interest points are identified using different detectors:

  1. Difference of Gaussian
  2. Hessian-Laplacian Detector

  For each of the detected keypoints we then compute a SIFT [7] descriptor using the VIREO system [3]. We use the K-means algorithm to cluster the descriptors from the training set into 500, 1,000 and 2,000 visual words. After quantization of the feature space, an image is represented by a histogram where the bins of this histogram count the visual words closest to image keypoints. We therefore obtain feature vectors of dimension 5,00, 1,000 and 2,000. This run uses only the 2,000 vocabulary.

- **Dense SIFT and ColorSIFT** We also use a dense sampling for the SIFT and ColorSIFT descriptors proposed by Koen Van de Sande [18]. We use their software provided in [1]. We created visual dictionaries of size 1,000, 4,000 and 10,000. We pool the quantized descriptors globally over the whole image. We also consider pooling according to a spatial pyramid (1, 2x2, 3x1), so that the corresponding feature vectors have a dimension 8 times the size of the dictionary. This run uses only the 10,000 vocabulary.

- **Saliency Moments descriptor** This is a holistic descriptor which embeds some locally-parsed information, namely the shape of the salient region, in a holistic representation of the scene, structurally similar to [11]. First, the saliency information is extracted at different resolutions using a spectral, light-weight algorithm [5]. The signals obtained are then sampled directly in the frequency domain, using a set of Gabor wavelets. Each of these samples, called "Saliency Components", is then interpreted as a probability distribution: the components are divided into subwindows and the first three moments are extracted, namely mean, standard deviation and skewness. The resulting signature vector is a 482-dimensional descriptor [13].

- **MEDA** We have proposed descriptors based on marginal distributions of the local descriptors. They have the advantage of a faster computation than bag-of-word construction, and have shown efficient performance. Those descriptors are described in [14].

- **ST-MPEG7** This is a spatio-temporal descriptor based on the temporal statistics of the MPEG-7 Edge Histogram descriptor.

Those feature vectors are quantized and then expanded using the Homogeneous Kernel Maps [19], which approximate the Histogram Intersection kernel by a scalar product. This allows to use linear SVMs classifiers on the expanded vectors, instead of SVMs with more complex kernels. The order of the Homogeneous Kernel Maps is 5.

To train the linear SVMs, we have implemented a variation of the PEGASOS algorithm [16], where we dynamically adapt the weight between positive and negative samples, as well as updating several models in parallel. For each concept, we select the best SVM parameters as those which maximize the Mean Average Precision (MAP) on a validation set.

We fuse the results of the classifiers using a linear SVM.

3 EURECOM Second Run: Run3

This run is an expansion of the previous one, where we add new descriptors by varying the size of the vocabularies used for the SIFT and Color-SIFT descriptors, as well as considering dense sampling without spatial pyramid. This provides a total of 28 descriptors, for which classifiers are trained. The resulting scores are fused using a linear SVM to provide the final score.

4 EURECOM Third Run: Run2

While the previous runs only involve the use of TRECVID training data, we included in this run classifiers that have been built on the ImageNet task, using some of the most advanced processing techniques. We used the CAFFE Deep Neural Net [6] developped by the Vision group of the University of Berkeley, for which both the source code and the trained parameter values have been made available. We also use Fisher Vectors combined with linear SVM [15], whose results on TREVVID have been made available through the French IRIM working group. For those two sets of classifiers, the training was performed on the ImageNet data only, and the classifiers have been applied unchanged on the TRECVID data. Each set contains 1000 classifiers for the 1000 ImageNet concepts, so the scores for each shot are accumulated in a 1000 dimension semantic feature vector, which is used as the description of the TRECVID shots, both for training and test data.

The rest of the processing is similar to the previous runs.

5 EURECOM Fourth Run: Run1

As introduced in the last two years, we also exploit the metadata provided with the TRECVID videos to benefit from the video uploader's information [8]. The TRECVID training and test data come from the same distribution and we have found that videos from several uploaders are distributed evenly over the corpus. We benefit from this information based on the assumption that an uploader is likely to upload videos with similar content. In other words most if not all videos uploaded by one user represent information that is not much different from one another. For example if a user runs a video blog about monuments in a certain city then almost all videos uploaded by that user will contain concepts like *sky* or *outdoor*. This information thus increases our confidence in the predictions of the concepts *sky* and *outdoor* if the test video is uploaded by the same user. This model is applied to selected concepts on top of Run 2.

The uploader model simply calculates the ratio of video shots uploaded by the uploader for each concept from the training data and modifies the output score of each new video shot if that video is uploaded by the same person. This uploader bias allows us to rerank the retrieval results. For each concept we calculate the probability of concept given uploader as:

$$p(c/u) = \frac{W_u^c}{|V_u|}$$

where V_u is the set of videos uploaded by uploader 'u' and W_u^c is the weightage of videos uploaded by uploader 'u' for the concept 'c'. This quantity:

$$W_u^c = \sum_{v \in V_u} \frac{|s \in v, s.t. s = c|}{|s \in v|}$$

is the sum of ratios of the number of shots labeled with concept 'c' to the total shots in that video for all the videos uploaded by 'u'.

We also calculate average uploader's probability for each concept as:

$$p(c) = \frac{W^c}{|V|}$$

where W^c is the total weightage of all the videos uploaded for concept 'c', given by:

$$W^c = \sum_u W_u^c$$

and V is the number of videos, or $V = \sum_u V_u$.

This model is computed on the training data separately for each concept. To apply the uploader's model to the test videos we calculate the coefficient α as:

$$\alpha(c,u) = \max\left(\frac{p(c/u) - p(c)}{p(c/u) + p(c)}, 0\right)$$

The score of each shot $P(c|s)$ from the previous run is modified in the following way:

$$p_u(c|s) = p_2(c|s) * (1 + \alpha(c,u))$$

6 Results Analysis

Run	MAP
Run1	0.2175
Run2	0.2025
Run3	0.1315
Run4	0.1151

Figure 2: Evaluation results for our (corrected) runs

In Figure 2, we indicate the performances (MAP) of our runs, as described in the previous chapter. Those figures are those provided by the manual evaluation performed by NIST. By comparing Run3 and Run4, we can see that considering several sizes of vocabularies allows a substantial increase in performance, probably due to a better reliability of the classifiers. The absolute performance of those classifiers remains however average.

When looking at the result of Run2, we can notice a large improvement in the performance. This means that the classifiers trained on ImageNet are able to provide very useful information to describe the contents of the TRECVID shots. Extra experiments that we have performed after the official TRECVID submissions lead to the conclusion that the Deep Network performs slightly better than the classifiers based on Fisher vectors.

Finally, the uploader model is able to use extra information and improve over the best results of Run3 to further increase the performance by more than 1

In figure 3 we display the comparative performance of the four runs on each of the concepts evaluated by NIST.

Because of a bug, our 2013 submissions were corrupted, therefore we do not have a possible comparison of these numbers.

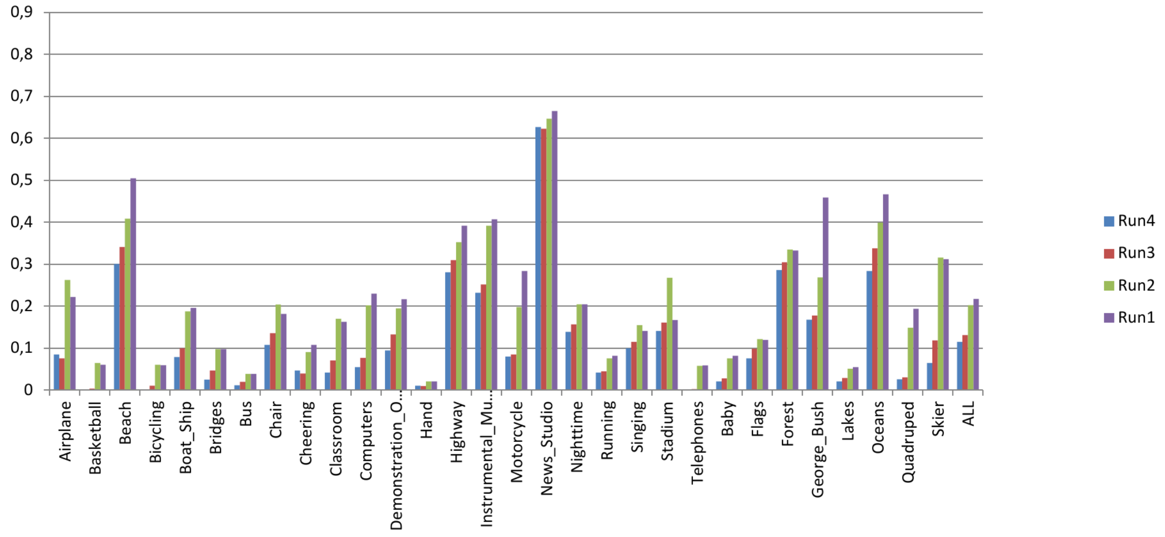

Figure 3: Results on the test set evaluated by NIST

7 Conclusions

This year EURECOM presented a set of systems for the Semantic INdexing Task. We included classifiers trained on ImageNet, and noticed that they provided very useful information for the description of TRECVID shots, resulting in a great improvement of the performance. As last year we confirmed that using uploader information allows to further improve the detection of certain concepts.

References

[1] Color descriptors, http://koen.me/research/colordescriptors/.

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

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

[4] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval (ECIR)*, pages 187–198, Glasgow, Scotland, mar 2008.

[5] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.

[6] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding, 2013.

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

[8] B. Merialdo and U. Niaz. Uploader models for video concept detection. In *CBMI 2014, 12th International Workshop on Content-Based Multimedia Indexing, 18-20 June 2014, Klagenfurt, Austria*, Klagenfurt, AUTRICHE, 06 2014.

[9] U. Niaz, M. Redi, C. Tanase, and B. Merialdo. EURECOM at TRECVID 2013: The light semantic indexing task. In *TRECVID 2013, 17th International Workshop on Video Retrieval Evaluation, 2013, National Institute of Standards and Technology, Gaithersburg, USA*, Gaithersburg, UNITED STATES, 11 2013.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971 –987, jul 2002.

[11] 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.

[12] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[13] M. Redi and B. Mérialdo. Saliency moments for image categorization. In *ICMR'11, 1st ACM International Conference on Multimedia Retrieval, April 17-20, 2011, Trento, Italy*, 04 2011.

[14] M. Redi and B. Merialdo. Direct modeling of image keypoints distribution through copula-based image signatures. In *ICMR 2013, ACM International Conference on Multimedia Retrieval, April 16-19, Dallas, Texas, USA*, Dallas, ÉTATS-UNIS, 04 2013.

[15] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *Int. J. Comput. Vis.*, 105(3):222–245, Dec. 2013.

[16] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 807–814, New York, NY, USA, 2007. ACM.

[17] C. Tanase and B. Mérialdo. Efficient spatio-temporal edge descriptor. In *MMM 2012, 18th International Conference on Multimedia Modeling, 4-6 January, 2012, Klagenfurt, Austria*, 01 2012.

[18] 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.

[19] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480 – 492, 2012.