EURECOM at TrecVid 2011: The Light Semantic Indexing Task

Usman Niaz, Miriam Redi, Claudiu Tanase, Bernard Merialdo,
Giovanna Farinella, Qian Li
Multimedia Department, EURECOM
Sophia Antipolis, France
{Usman.Niaz,Miriam.Redi,Claudiu.Tanase,Bernard.Merialdo,
Giovanna.Farinella,Qian.Li}@eurecom.fr

November 7, 2011

1 Abstract

This year EURECOM participated in the TRECVID light Semantic Indexing (SIN) Task for the submission of four different runs for 50 concepts. Our submission builds on the runs submitted last year at the 2010 SIN task by adding more effective visual features to the third system built last year, the details of which can be found in [10]. Two of our systems target specific objects based detection.

Our basic run adds densely extracted SIFT features to the pool of features of last year's basic run. The dense SIFT proves to be effective for concepts such as "Nighttime" accounting for a very low number of keypoints when extracted using a conventional log or hessian based detector. Then according to the third run from last year we add textual metadata based information that has been provided with the 2011 video database to the visual features. We improve the retrieval task by adding two more global descriptors to visual features with one capturing temporal statistics along a sequence of shots and the other capturing salient details or gist of an image. Further we enhance the visual recognition of some semantic concepts based on the detection of local objects like computer screens or scene text, and human detection like Male or Female persons.

The runs are composed as follows:

1. **EURECOM_Fusebase** This run fuses a pool of visual features, namely the Sift [7] descriptor extracted through dense, log and hessian methods, the Color Moments global descriptor, the Wavelet Feature and the Edge Histogram. On top of this, the information mined from the textual metadata files related to each video is added.

2. **EURECOM_Videosense_SM** This run adds Spatio-Temporal [12] and Saliency Moments feature [9] to the visual features pool of the previous run before adding the textual information.

3. **EURECOM_OS** This run adds to the previous run some local object detectors for 19 concepts.

4. **EURECOM_HS** This run adds to the runs 3 human specific detectors for 8 concepts like for example "Male_Person", "Old_People" etc..

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 of two main modules: first, a model is built by combining a pool of visual features,which is then extended by adding textual feature based on the video metadata.

1. ***Visual Feature Extraction and Fusion***: In this stage, 6 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:

   - **Bag of Words with SIFT** Three sets of interest points are identified using different detectors:

     (a) Saliency Points
     - Difference of Gaussian
     - Hessian-Laplacian Detector

     Each of these key points is then described with the SIFT [7] descriptor, using the VIREO system [1].

     (b) Dense extraction
     This case differs from the previous two saliency points detectors as SIFT features are extracted at points described by a predefined grid on the image [11]. The points on the grid are distanced 8-pixels apart.

     For each of these three extracted SIFT features, a Bag of Words (BoW) or visual vocabulary is built through quantization. We use K-means algorithm to cluster the descriptors from training set into 500 visual words based on the experiments on the development set. 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.

- **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.

A one vs all Support Vector Machine (SVM) (see implementation details in [2]) is trained for each feature. For each concept c a model based on each feature extracted from the training data is built and for each SVM classifier the value of the parameters are selected through exhaustive grid search by maximizing the Mean Average Precision (MAP) on a validation set. Such model is then used to detect the presence of c in a new sample s based on each feature.

We obtain thus, for each concept c and keyframe s, 6 feature-specific outputs that we call $p_n(c|s)$, $n = 1, \ldots, 6$. 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 appearance of the concept c in the keyframe s given the set of visual features. The dense SIFT feature dominates the weight distribution for most of the concepts during the fusion.

2. ***Term Frequency Metadata***: a textual feature module is added to the previous visual-only feature pool after fusion.

   Since last year, a set of XML-formatted metadata is available with each video, containing a textual description of the video context. 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 word that appears in the metadata file of video 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 visual features part of this run, obtaining, for each shot $s \in v$,

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

   The value β is estimated on the development data.

This step was performed only for the concepts for which adding this module was introducing a significant improvement in the final MAP (in the development stage).

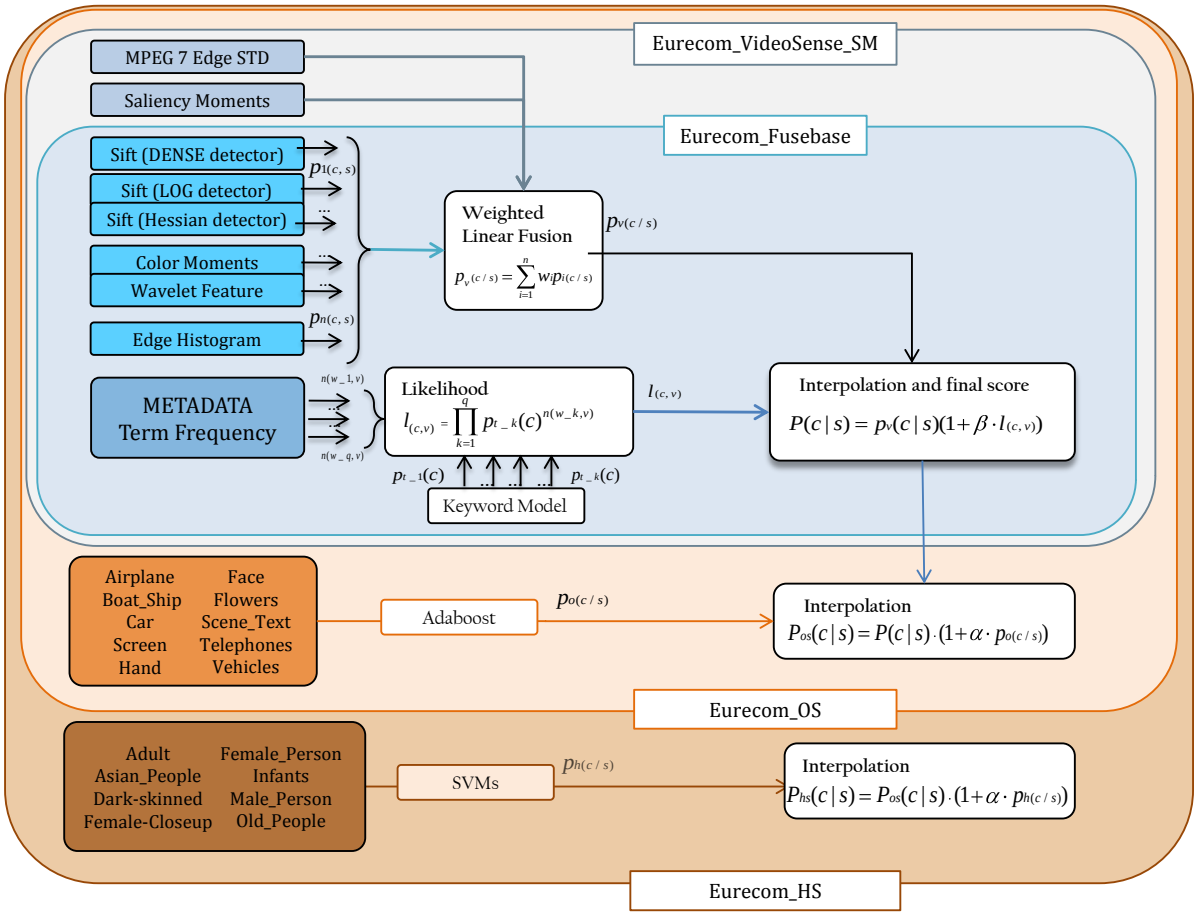

Figure 1: Framework of our system for the semantic indexing task

3 EURECOM Second Run: EURECOM_Videosense_SM

In this run, the Visual Feature Extraction and Fusion module of the basic run is improved by adding two new descriptors to the visual feature pool, namely the Spatial Temporal descriptor [12] and the Saliency Moments descriptor [9]. Then, similar to the previous run, the textual feature based on metadata is combined with the visual feature pool now containing 8 features. Following is a little description of each of these two new visual features.

- **ST-MP7EH Feature** We add to the visual feature pool of run 1 an efficient temporal statistic based global descriptor that is sampled on equally-spaced frames in the video [12]. The descriptor, called ST-MP7EH is simple, fast, accurate, has a very low memory footprint and works surprisingly well in conjunction with the existing visual descriptors. The ST-MP7EH descriptor detects the evolution in time of visual texture by computing the (2D) MPEG-7 edge histogram for each frame of the analyzed video giving an 80 value feature vector. This is done for N contiguous frames, with a $frameskip$ of 4 to reduce computation, resulting in an $N \times 80$ matrix. For each column of this matrix average and standard deviation is calculated which gives it a fixed dimension of 160. The values are of

the same order of magnitude as the ones from the image descriptor. The spatial information captured by the image descriptor is conserved by means of average and standard deviation, and important temporal relationships are established with the presence of the standard deviation. We use the 160 dimension spatial temporal descriptor to train an SVM classifier for embedding it with other visual descriptors.

- **Saliency Moments Feature** Additionally, a holistic descriptor for image recognition, namely the Saliency Moments feature [9] is added to run 1. SM embeds some locally-parsed information, namely the shape of the salient region, in a holistic representation of the scene, structurally similar to [8]. 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 that we use as input for traditional support vector machines and then combine with the contributions of the other visual features.

For the two new visual features the SVM parameters are trained via grid search and weighted linear fusion is used to combine the outputs $p_n(c,s)$, $n = 1, \ldots, 8$ of eight visual features into a single $p_v(c,s)$ for each concept c in the keyframe s. This output is then fused with the textual metadata feature as done in the previous run.

4 EURECOM Third Run: EURECOM_OS

In this run we focus on local object detectors motivated by the fact that many concepts in this year's light-SIN task depict objects that can be easily differenciated from the environment or background.

We follow Viola and Jones' object detectors framework [13] using a cascade of boosted classifiers for rapid object detection based on the openCV haar training API. We train boosted classifiers on 8 objects (see Figure 1 for a list), that are semantically related to a total of 15 concepts in the set of 50 semantic concepts for the light-SIN task. An example is that the object "boat" is present in the concept "Boat_Ship" and it should also yield positive for the concept "Waterscape_Waterfront". Additionally, we add a face specific detector similar to the first run of [10] to the concepts for which adding it showed significant improvement on the development data.

The output $P(c|s)$ from run 2 is interpolated with the output of the local object detector (face or otherwise), defined as $p_o(c|s)$, ran on the keyframe s, (see Figure 1). The interpolation is computed as follows: $P_{os}(c|s) = P(c|s)(1 + \alpha \cdot p_o(c|s))$, where the parameter α is estimated by maximizing the MAP of the development set.

5 EURECOM Fourth Run: EURECOM_HS

In this run human concepts detection is improved by using Histogram of Oriented Gradients (HOG) to describe the face region. Eight concepts are evaluated during this run: Adult, Asian People, Dark skinned People, Female Person, Female Human Face Close Up, Infants, Male Person and Old People, by adding the descriptor on top of run 3 for these concepts only.

HOG has been proposed by Dalal and Triggs for pedestrian detection in static images [3]. It is broadly used in computer vision, especially in the field of object detection. HOG has been recently used by Guo et al. [4] for gender classification purpose and it is shown to be very effective on benchmark face databases for both gender and ethnicity classification.

For all the concepts, with exception of Female Human Face Close Up, the following protocol is used. First of all, faces are detected using the algorithm provided by W. Kienzle [6]. Furthermore, face region is cropped and resized in order to be 100x100. Finally, HOG over cropped images is evaluated using 3x3 cell blocks of 6x6 pixel cells with 9 histogram channels. As for the other visual features, the extracted descriptor for each face detected is used as input for an SVM.

For Female Human Face Close Up, we use as information the dimension of the face region returned by the face detector: given the Female Person confidence scores, Female Human Face Close Up scores are obtained setting to zero the ones in which the face region is less than half of height or width of a frame.

For the eight concepts containing human objects the probability $p_h(c|s)$ of the presence of the concept c in the keyframe s is interpolated with the output $P_{os}(c|s)$ of run 3, (see Figure 1). This is done similarly as done for the run 3 giving $P_{hs}(c|s) = P_{os}(c|s)(1 + \alpha \cdot p_h(c|s))$ where α is estimated by maximizing the MAP of the development set.

6 Results Analysis

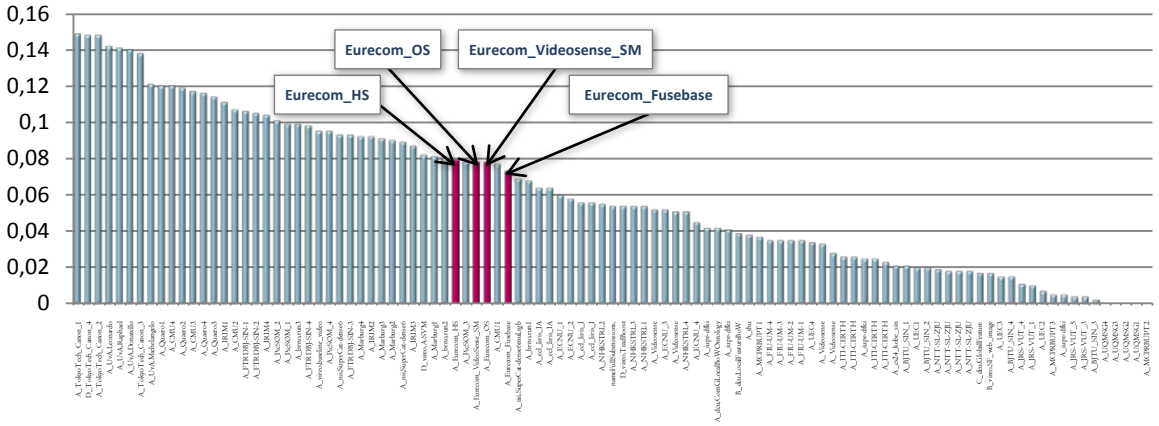

Figure 2: Evaluation results for our submitted runs

In Figure 2, the performances (MAP) of the various systems submitted for the light-SIN task are presented. The evaluations were carried out for 23 out of 50 concepts in the light-SIN task.

For further evaluation of our results on the 23 concepts we have separated our concepts into two sets namely (*i*) the concepts that (tend to) contain local objects for which all four runs including local and human specific object detectors were submitted, results in Figure 4, and (*ii*) other concepts for which only run 1 and 2 containing visual and textual descriptors were performed, shown in Figure 3.

The first two runs are based on classical visual features and textual information, as mentioned in Sec. 2-3. Shown in Figures 3 and 4 are the concept-specific final performances. Figures 3 and 4 also show the expected MAP (i.e. the value expected from the training phase) for all the runs. Compared to the estimated Mean Average Precision, the performances on the test set decrease of about 58% for the basic run and 55% for the Eurecom_Videosense_SM. 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. For concepts that are depicted in a sequence of keyframes we note that the added spatio-temporal information along with the saliency moments descriptor strongly impacts the performance with the concepts "Dancing" and "Walking" giving a performance even better than that acquired in the training phase. For all the concepts in Figure 3 the MAP does not improve beyond run 2 as local detectors are not evaluated on these concepts.

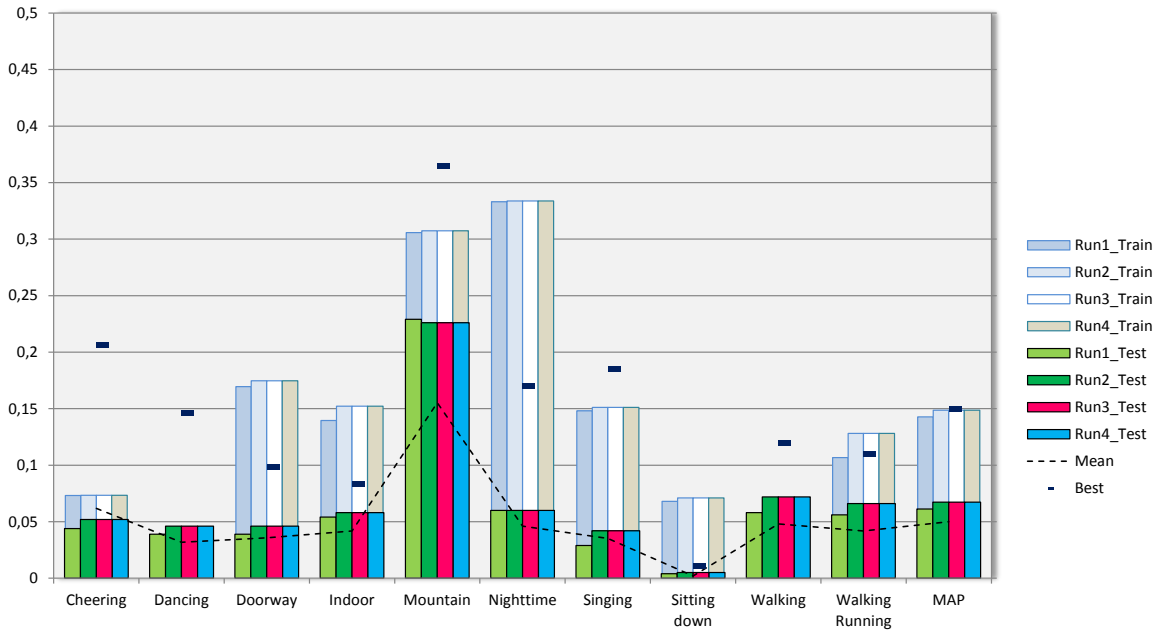

Figure 3: Development and test results comparison for concepts worked on mainly by run 1 and 2

Similar increase in performance is seen for most of the concepts in Figure 4 from run 1 to run 2 for the training and the test evaluations alike. Here the development projections are not

too optimistic as we see a performance difference of only about 26% between the training and the test assessments. Though the boosted local object detectors always increase performance on the training set it actually degrades test results in run 3 for some concepts.

"News Studio" for which the retrieval performance is already high till run 2 is further improved by Face and Screen local detectors in run 3. While for "Adult" and "Old People" face detector degrades performance in run 3 while the human specific detector of run 4 seems to work better. Similarly in run 3 addition of Face detector to the concept "Explosion Fire" worsens the performance which is then carried on to the run 4 as it extends the run 3.

Human specific detectors works well for concepts "Female Person" and "Female Human Face Closeup" contrary to "Male Person".

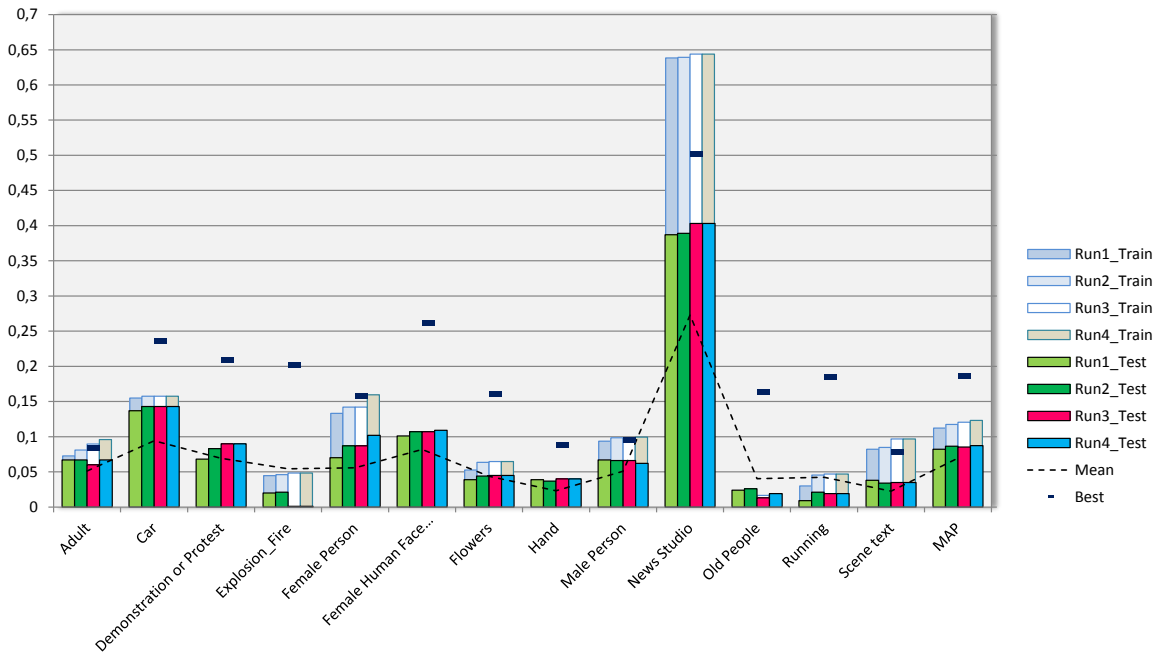

Figure 4: Development and test results comparison for concepts with local objects, run 1-4

7 Conclusions

This year EURECOM presented a set of systems for the light Semantic Indexing Task. As last year we confirmed that adding textual features extracted from the metadata improves the visual-only based systems. Spatial-temporal statistics based descriptor improves performance on concepts that are spread through a sequence of keyframes. Saliency distribution is shown to provide complementary information with respect to traditional visual features, improving the final AP for global concepts like "Indoor".

The boosted local object detectors generally improve indexing task for specific concepts but the face detector from our last year's submission degrades performance for most of the concepts

where it is used. Human specific features also prove to be effective though their impact is only on a small subset of concepts.

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] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 - Volume 01*, pages 886–893, Washington, DC, USA, 2005.

[4] G. Guo, C. Dyer, Y. Fu, and T. Huang. Is gender recognition affected by age? In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 2032 –2039, October 2009.

[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] W. Kienzle. http://pmsol3.wordpress.com, 2008.

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

[8] 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.

[9] 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.

[10] M. Redi, B. Mérialdo, and F. Wang. EURECOM and ECNU at TrecVid 2010 : The semantic indexing task. In *TRECVid'2010, 14th International Workshop on Video Retrieval Evaluation, November 15-17, 2010 National Institute of Standards and Technology, Gaithersburg, Maryland USA*, 11 2010.

[11] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77:157–173, May 2008.

[12] 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.

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:511, 2001.