EURECOM at TrecVid 2012: The Light Semantic Indexing Task

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

October 29, 2012

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 2011 SIN task with the first two runs following the same pattern as those of last year. The details of 2011 system can be found in [8]. One of our run adds uploaders bias to the pool of visual features while another run is prepared in collaboration with ECNU.

Our basic run adds visual features based on larger vectors to the pool of features of last year's base run. Larger dictionaries provide a finer representation of the visual/clustering space and increase the precision of the retrieval task. Like in last year's submission we add two global descriptors to visual features with one capturing temporal statistics along each shot and the other capturing salient details or gist of a keyframe. Then we add textual metadata based information that has been provided with the 2012 video database to the visual features. We further benefit from the metadata by including uploaders bias to increase scores of videos uploaded by same users.

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, the Edge Histogram, the texture based Local Binary Pattern feature and the dense color pyramid [14].

2. **EURECOM_Videosense_SM** This run adds Spatio-Temporal [13] and Saliency Moments feature [11] to the visual features pool of the previous run. On top of this, the information mined from the textual metadata files related to each video is added.

3. **EURECOM_uploader** This run adds uploaders' information to the previous run.

4. **EURECOM_ECNU** This run combines EURECOM_Videosense_SM visual run with visual features from ECNU.

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

The remainder of this paper briefly describes the content of each run (Sec 2-5), including feature extraction, fusion and reranking 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: EURECOM_Fusebase

This run comprises a number of visual features ranging from local features like gathering information from image pixels and local binary patters to global image descriptions like wavelet and bag of words histograms.

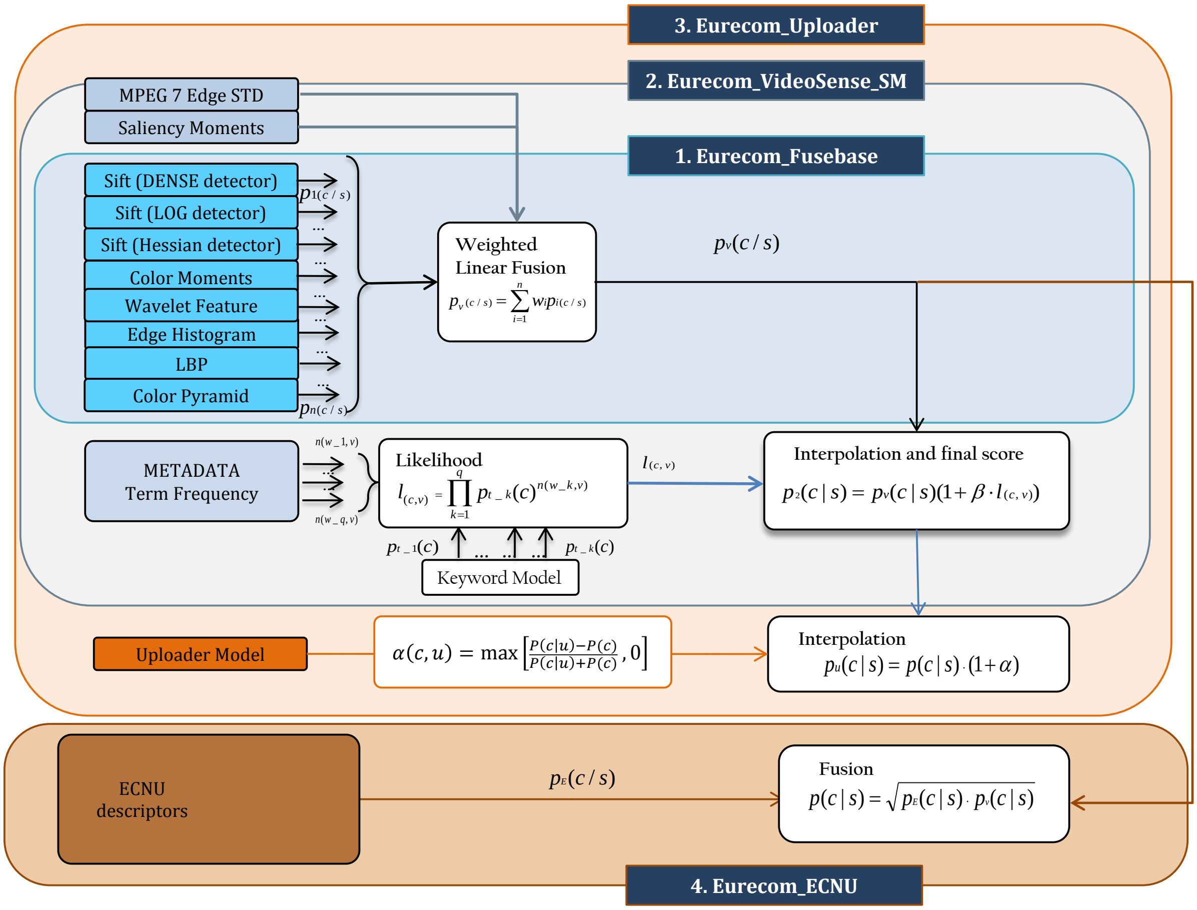

Figure 1: Framework of our system for the semantic indexing task

In this stage, 8 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:

    1. 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 [3].

    2. 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 [12]. 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 and 1000 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.

- **Local Binary Pattern (LBP)** Local binary pattern describes the local texture information around each point [9], 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.

- **Color Pyramid** Color pyramid is the spatial pyramid based representation of the dense color SIFT descriptor provided by UVA [14]. We use their software provided in [1].

A one vs all Support Vector Machine (SVM) 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.

For SIFT features, after performing experiments on the validation set we selected dictionary size of 1000 for log and hessian based descriptors and dictionary of 500 words for the densely extracted descriptors. Further for the three SIFT feautures we train a non linear SVM with chi-square kernel of degree 2 [4], while for rest of the five visual features a linear SVM is used to learn from a suitable feature map (homogeneous kernel map) built by the histogram intersection kernel [15].

We obtain thus, for each concept c and keyframe s, 8 feature-specific outputs that we call $p_n(c|s)$, $n = 1, \ldots, 8$. 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 and the color pyramid feature dominate the weight distribution for most of the concepts during the fusion.

3 EURECOM Second Run: EURECOM_Videosense_SM

This run is composed of two modules: first, as done in the previous year [8], 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 [13] and the Saliency Moments descriptor [11]. Then textual feature based on the video metadata is added to the visual features pool. This metadata is provided with the trecvid 2012 collection as in the previous years. The two steps performed in this run are detailed below.

1. ***Visual Feature Extraction and Fusion***: Following is a summarized description of each of the two new visual features added to the first run:

    - **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 [13]. 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 [11] 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 [10]. First, the saliency information is extracted at different resolutions using a spectral, light-weight algorithm [6]. 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 $p_n(c, s)$, $n = 1, \dots, 8$ 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.

2. ***Term Frequency Metadata***: A textual feature module is added to the visual-only feature pool (Run 1 + 2 global descriptors) after fusion. The process is similar to that followed in our previous submissions.

   Since 2010, 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 $s \in v$,

$$p_2(c|s) = p_v(c|s)(1 + \beta \cdot l(c, v))$$

   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).

4 EURECOM Third Run: EURECOM_uploader

In this run we explore the metadata provided with the TRECVID videos to benefit from the video uploader's information. The trecvid training and test data comes from the same distribution and we have found that videos from several uploaders are distributed evenly over the corpus. We benefit from this information keeping in mind that an uploader is likely to upload similar videos. 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 certain 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:

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

Cross validation on the development set showed that the uploader model was efficient for 25 out of 50 concepts. It was therefore applied to those 25 concepts for the final run and the other 25 were left unchanged.

5 EURECOM Fourth Run: EURECOM_ECNU

This run combines visual features form Eurecom and ECNU, china. EURECOM_Videosense_SM comprising 10 visual features is combined with indigenous visual features from ECNU. Details of their descriptors and their implementations can be found in their notebook paper [5]. They also provide a probability score $p_E(c|s)$ for each keyframe which is fused with our score using a simple linear fusion.

$$p(c|s) = \sqrt{p_E(c|s) * p_v(c|s)}$$

6 Results Analysis

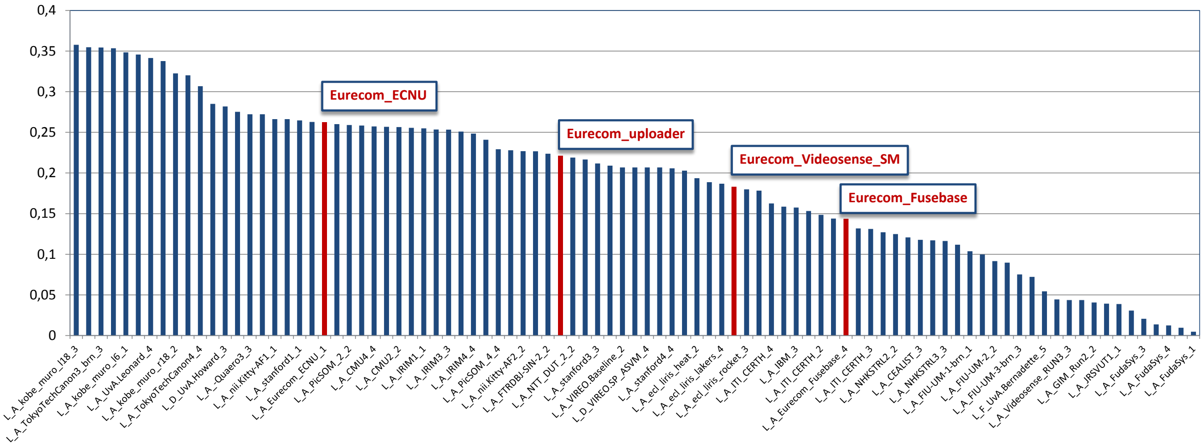

Figure 2: Evaluation results for our submitted runs

In Figure 2, the performances (MAP) of the various systems submitted for the light-SIN task are presented. This year 15 out of 50 concepts were evaluated for this light task. We can see that the performance of our runs vary significantly with our best two runs acquiring positions 21 and 42.

Figure 3 shows performance of our four runs on the TRECVID 2012 developpment set. For cross validation we divided the 2012_b set randomly into 3 partitions. We calculated average precisions (on top 2000 shots) for each of these three partitions to get 3 different scores for each each concept and then averaged those scores. We then find the maximum average score to choose best classifier parameters. For all the 50 concepts in the light SIN task our MAP scores on the average of average precisions are 0.169 for *Eurecom_Fusebase*, 0.179 for *Eurecom_Videosense_SM*, 0.211 for *Eurecom_uploader* and 0.202 for *Eurecom_ECNU*. Concerning results on the whole training set as well as on the subset of 15 concepts evaluated by NIST, Run 3 with the uploader model outperforms all the other Runs.

In figure 4 the results evaluated by NIST are shown. The fact that results on test set are better than those acheived on the training set may be explained by the fact that we use the

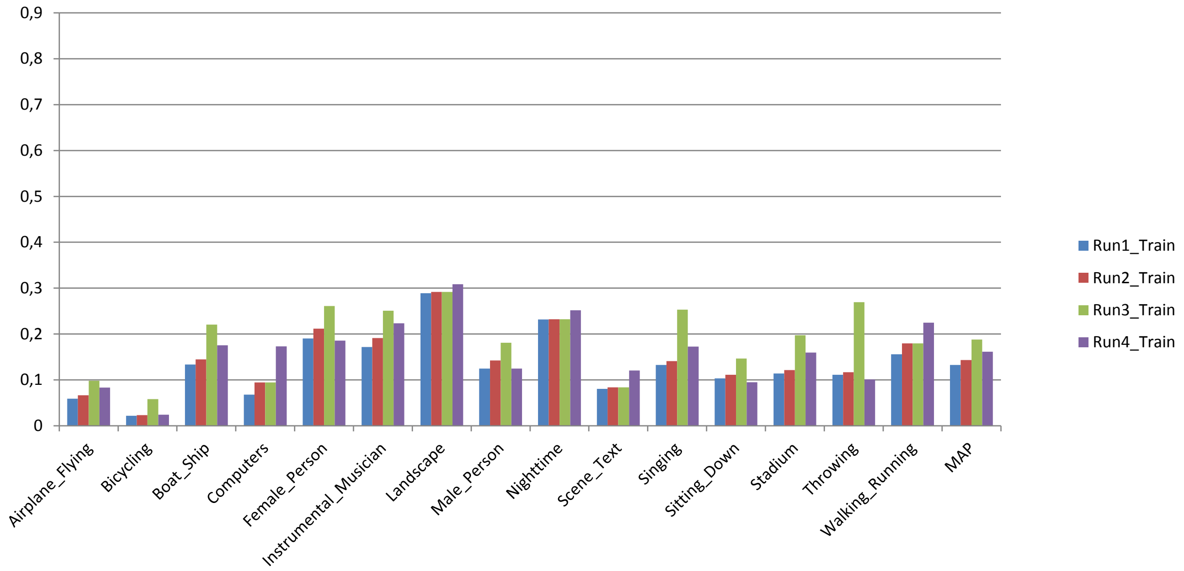

Figure 3: Results on the development set for the 15 concepts evaluated by NIST

average of the average precisions on three subsets of the validation part of the developpment set.

The first two runs are based on classical visual features, as mentioned in Sec. 2-3, with the second run including a reranking based on text features. The second run improves on almost all the evaluated concepts with considerable performance gain for *Boat_ship* due to the saliency moments descriptor and for *Airplane_Flying*, *Throwing* and *Walking_Running* mainly due to the presence of spatial temporal descriptors.

Run 3 improves further on Run 2. The uploader model improves average precision for every concept to which it is applied except for *Male_Person* where there is a negligible drop in the already hight score. Using uploader's information to detect concepts proves beneficial here as it increases score even when the visual descriptors failed to retrieve the concept *Bicycling*.

ECNU has provided strong visual features that improves retrieval score for most of the evaluated concepts. ECNU features are added to the Run 2 and the MAP for 15 evaluated concepts has increased 43%. The improvement over Run 3 is 18%. This Run is placed at the 21st position among all the submissions for the 2012 light semantic indexing task.

7 Conclusions

This year EURECOM presented a set of systems for the light Semantic Indexing Task. As last years 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

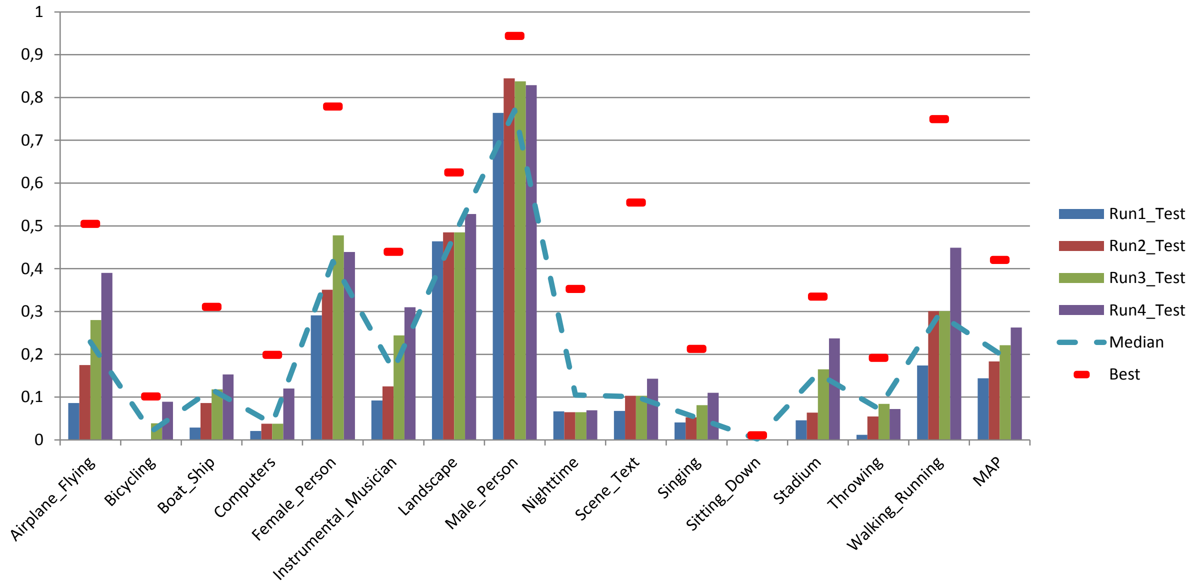

Figure 4: Results on the test set evaluated by NIST

final AP for global concepts.

Information based on uploader of the video tells alot about the video as users tend to upload similar videos. This phenomenon is reflected in the results with the use of a simple reranking model based on videos' uploader. Visual features provided by ECNU when combined with our visual features makes a strong vidual classifier that outperforms the performance of all the runs submitted by EURECOM and ECNU independently.

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

[5] D. Z. F. Wang, Z. Sun and C. W. Ngo. Semantic indexing and multimedia event detection: Ecnu at trecvid 2012. *TRECVID 2012, 15th International Workshop on Video Retrieval Evaluation, National Institute of Standards and Technology, Gaithersburg, USA*, 2012.

[6] 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.

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

[8] U. Niaz, M. Redi, C. Tanase, B. Merialdo, G. Farinella, and Q. Li. EURECOM at TRECVID 2011: The light semantic indexing task. In *TRECVID 2011, 15th International Workshop on Video Retrieval Evaluation, 2011, National Institute of Standards and Technology, Gaithersburg, USA*, Gaithersburg, UNITED STATES, 11 2011.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.