Semantic Indexing and Multimedia Event Detection: ECNU at TRECVID 2012

Feng Wang[†], Zhanhu Sun[†], Daran Zhang[†], Chong-Wah Ngo[‡]

[†] *Dept. of Computer Science and Technology, East China Normal University*

[‡] *Dept. of Computer Science, City University of Hong Kong*

1 Abstract

This year we participated in two tasks: Semantic Indexing (SIN) and Multimedia Event Detection (MED). In this paper, we present our approaches and discuss the evaluation results.

Semantic Indexing (SIN): For video semantic indexing, we focus on the performance improvement by using a Weighted Hamming Embedding kernel compared with traditional BoW approaches. Below are the brief descriptions of our submitted runs.

- L_A_ECNU_4: This run serves as our baseline. Color, texture, audio and local features are used for detector training.
- L_A_ECNU_3: We employ Hamming Embedding to improve the performance of traditional BoW approaches with SIFT feature.
- L_A_ECNU_2: Soft weighting is integrated with Hamming Embedding.
- L_A_ECNU_1: The approach proposed in our previous work [13] is employed to weight the informativeness of visual words.

Multimedia Event Detection (MED): In this year's MED task, we revise and validate our approaches proposed in [12] for event detection on the large video corpus. The following runs are submitted.

- PS_EKFull_AutoEAG_c-run3: Detectors trained using staic visual, audio and semantic features.
- PS_EKFull_AutoEAG_c-run2: Relative motion histograms are calculated between visual words and used for event detection.
- PS_EKFull_AutoEAG_p-baseline: Feature selection is employed to address the efficiency issue.

2 Semantic Indexing

2.1 Baseline

In this year's semantic indexing task, we adopt our approach in TRECVID 2009 [10] as the baseline system. Five keyframes are extracted from each shot. The following features are extracted.

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 [3].

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

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.

SIFT: Difference of Gaussian (DoG) [1] and Hessian Affine [2] detectors are used to detect local interest points, and 128-dimension SIFT feature [1] is extracted to describe each local image patch. A visual vocabulary is generated by clustering the SIFT descriptors with k-means algorithm. Each SIFT descriptor is then mapped to the nearest visual word to form the BoW histogram for each keyframe.

ColorSIFT: Compared with SIFT extracted from grayscale images, ColorSIFT [20] captures the color information and is also employed in our work. One 128-dimension vector is obtained for each of 3 channels.

Local Binary Pattern (LBP): Different from SIFT features extracted from the locally salient keypoint, local binary pattern describes the local texture information around each point [4]. We employ the implementation in [6] 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): A 80-dimension feature is calculated as the average and variance of MFCC coefficients.

For classification, SVM with RBF kernel [5] is adopted. We use χ^2 distance for SIFT, ColorSIFT, and LBP features, while Euclidean distance for others. Due to the limitations of computing resources, we only submit results for the light task. To fully make use of concept relations, we train detectors for another 40 concepts. The additional concepts are selected based on their semantic relations to the required ones. For each required concept c_1, the detection scores are then increased (or decreased) according to the scores of another concept c_2 that implies (or excludes) c_1 by a factor f, where the parameter f is learned from the development set.

2.2 Weighted Hamming Embedding Kernel

In this year's SIN task, we focus on validating the performance improvement gained by our proposed Weighted Hamming Embedding kernel. Our approach addresses two problems existing in the widely used BoW approach.

The first problem with BoW approach is the information loss caused by the quantization of SIFT space during visual vocabulary construction. In previous works, soft weighting [14, 15, 16] is proposed to alleviate this problem by assigning each descriptor to few neighboring visual words. Although this ambiguous approach can partially alleviate the information loss problem, all the descriptors assigned to a word are assumed to be idential no matter how different they are from each other. In our approach, we employ Hamming Embedding to alleviate this problem. During quantization stage, each descriptor is associated with a binary signature [11] which encodes its location information inside the Voronoi cell corresponding to the assigned visual word. The Hamming distances between different descriptors in the same cell are then calculated to approximate the Euclidean distance between them. Compared with the ambiguous assignment in soft weighting, our approach estimates the distance between different image samples with higher precision. In practice, Hamming Embedding and Soft Weighting can be integrated. The results are reported in our run L_A_ECNU_2. Each descriptor is first mapped to the 3 nearest visual words. The above proposed Hamming Embedding is then employed for the distance measure between different descriptors.

Secondly, we employ our work in [13] to weight the informativeness of visual words for different concepts. Given a concept, usually only some of the visual words frequently appear while the presence of the others are nearly random. In other words, some visual words are more informative or important for the detection of a specific concept, while the others may be noisy. In most current approaches, for different concepts, each visual word is treated equally. The discriminative ability of those informative visual words would be seriously reduced. To cope with this problem, for the detection of each concept, we propose to weight the informativeness of different visual words. This is carried out in the framework of kernel optimization. Kernel Alignment Score [22] is employed to evaluate the discriminative ability of an SVM kernel. The weights of different words are estimated by maximizing the KAS score. Finally, the more important visual words are assigned with larger weights in the resulting SVM kernel and contribute more to the detection of the specific concept.

2.3 Evaluation Results

Figure 1 shows the performances of our submitted runs among all submissions, and Figure 2 shows the per-concept results returned by NIST. Compared with the baseline (L_A_ECNU_4), Hamming Embedding (L_A_ECNU_3) improves the MAP (Mean Average Precision) by 14.80%. Among all the evaluated concepts, significant improvement can be observed for the concepts *Airplane_Flying* (20.0%), *Bicycling* (38.5%), *Boat_Ship* (23.9%), *Computers* (46.4%), *Nighttime* (36.4%) and *Instrument_Musician* (29.4%). The improvement shows to be consistent for all concepts. This demonstrates the effectiveness of our approach in alleviating the information loss problem of BoW based approach. By further incorporating soft weighting (L_A_ECNU_2), another 4.65% improvement is achieved. This improvement is basically consistent with the results reported in [16]. This shows that Hamming Embedding and Soft Weighting could be complementary to each other.

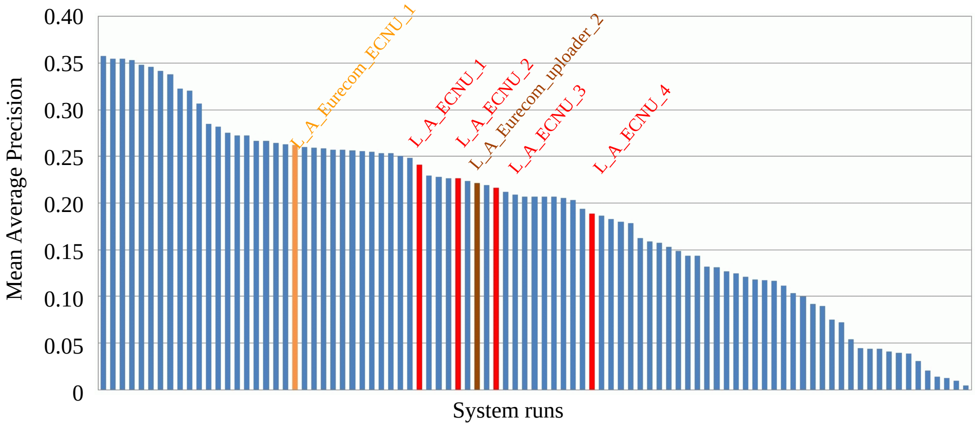

Figure 1: Performance of our submitted runs among all systems for SIN task. The red columns correspond to our submitted runs.

By weighting the informativeness of visual words, an improvement of 6.26% is gained. Among all concepts, significant improvement could be observed for *Airplane_Flying* (49.1%), *Computers* (11%), *Boat_Ship* (8.9%), and *Nighttime* (8.5%). Noting that the proposed approach is compared with the whole system (instead of only BoW based approach), this improvement could be considered significant.

Figure 1 also shows a collaboration run (L_A_Eurecom_ECNU_1) by fusing the best runs from ECNU (L_A_ECNU_1) and Eurecom (L_A_Eurecom_uploader_2) respectively to demonstrate the effectiveness of system combination. The details of this run could be found in [21].

3 Multimedia Event Detection

This is the first time for us to participate in the MED task. We basically employ the system design and features that have been proven effective by other teams in MED10 and MED11 [18, 17]. In addition, we revise and validate the performance of the motion features proposed in our previous work [12].

3.1 Feature Extraction and Classifier Learning for the Basis Run

In this work, we first build a basis system (c-run3) by employing the existing features and classifiers. Below is a brief description.

For static features, one keyframe is sampled in every 5 seconds along each video sequence. Features are then extracted in sampled frames. Local interest points are detected using DoG (Difference of Gaussian) [1] and Hessian Affine [2] detectors. 128-dimension SIFT descriptors [1] are employed to describe the local image patches. ColorSIFT [20] is further used to employ the missing color information in SIFT. Besides low-level features, we extract mid-level features,

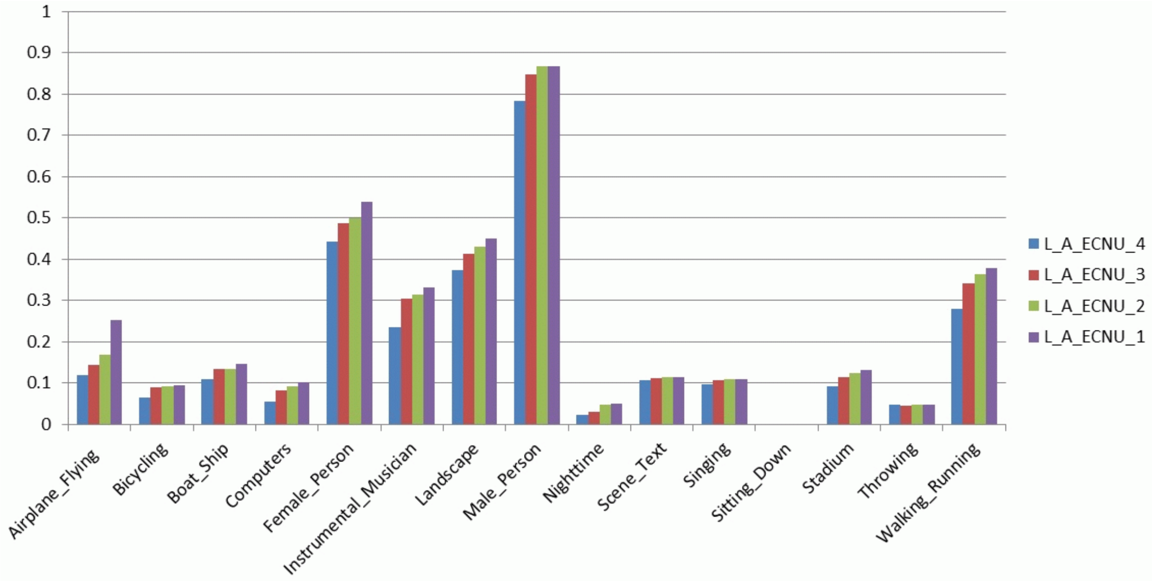

Figure 2: The performances of our submitted runs for the concepts evaluated by SIN 12.

i.e. concept scores for event detection. Due to the lack of annotations on MED development set, we simply borrow the concept detectors from SIN task. In total, 46 concepts which are semantically related the events are manually selected. The scores output by the corresponding detectors are used as the semantic features. Besides visual features, audio information is also considered in our system. MFCC coefficients which are widely in previous works are extracted in every audio frame of 50ms, where each frame overlaps with neighboring ones by 25ms.

For SIFT, ColorSIFT and MFCC, Bag-of-Words approach is employed in feature representation. The feature spaces are first quantized into 2000, 4000, 8000 words respectively. Soft weighting is used for word assignment. Each descriptor is mapped to the 3 nearest words.

For classifer learning, LIBSVM [5] is employed. We consider two kinds of approaches: $\chi^2 - RBF$ kernel and EMD (Earth Mover's Distance) based temporal matching [19]. $\chi^2 - RBF$ kernel is employed for SIFT, ColorSIFT and MFCC features, while temporal matching for SIFT, ColorSIFT and concept scores. The results of all classifiers are combined with linear weighted fusion. In the submission, the thresholds are determined by minimizing the NDC scores.

3.2 Relative Motion Histogram

Besides the above features, we employ the approaches proposed in our previous work [12]. We mainly validate the performance improvement by employing motion features. For each video clip, we extract motion histograms on visual words and relative motion hitograms between different visual words. Keypoints are first detected and tracked between neighboring frames. The motion histograms are calculated as

$$m_i(v_a) = \sum_{p \in N_{v_a}} D_i(m_p) \tag{1}$$

where N_{v_a} denotes the set of points mapped to word v_a, m_p is the motion vector of point p, and $D_i(\cdot)$ projects m_p to the $i-th$ direction. The relative motion between two words v_a, v_b is calculated as the sum of motion between all keypoints mapped to them respectively.

$$r_i(v_a, v_b) = \sum_{p \in N_{v_a}, q \in N_{v_b}} D_i(m_p - m_q). \tag{2}$$

where $m_p - m_q$ is the relative motion vector between two keypoints p and q. To alleviate the ambiguity in visual word assignment, the motion is expanded to the nearest 3 words. The details can be found in [12]. For the histogram calculation, two kinds of video volumes of different lengths are investigated: 5-second volumes and the whole video sequence. During classifer learning, EMD based temporal alignment is used for the former, and $\chi^2 - RBF$ kernel is used for the latter.

One problem with the above feature is the high dimension which brings heavy computation load to feature extraction and classification process. In this work, we alleviate this problem by selecting the motion features which are the most effective for event detection. A specific event is usually related to only few objects (or concepts) corresponding to a small set of visual words and limited motion patterns between them. As the result, most elements in the above motion histogram are useless (even noisy) in describing and detecting the event. Based on this observation, we reduce the feature dimension by selecting only those informative motion features. This is achieved by employing an approach similar to our work in [13]. To speed up the process of weight estimation, a stepwise gradient based approach is adopted. The weight of each element can only be 1, $\frac{3}{4}$, $\frac{2}{4}$, $\frac{1}{4}$, or 0. The motion features which are the most informative in detecting a specific event are selected. The visual words and motion features that are less informative (with a weight of $\frac{1}{4}$ or 0) are not examined during feature extraction and SVM learning. Thus, the dimension of the resulting feature is significantly reduced. At the same time, the performance is not affected since the removed features are mostly useless for event detection.

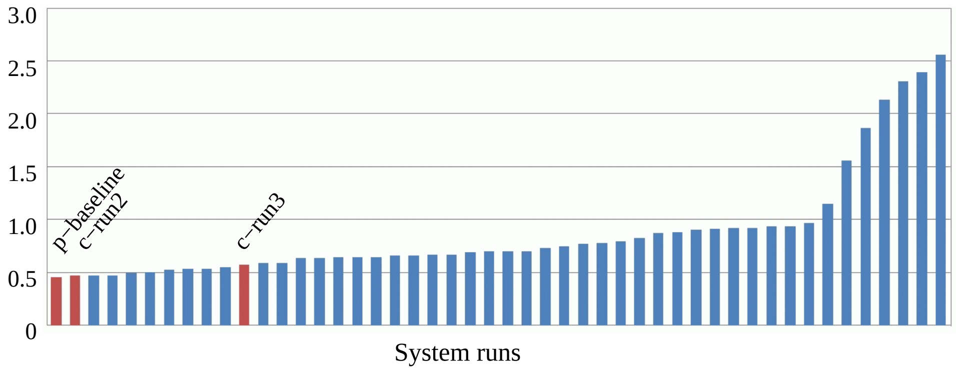

Figure 3: Performance of our runs on MED task. To better compare different systems, the last run with the highest NDC (9.30) is not shown in this figure.

Runs	Number of Events Meeting Goals	
	Actual Decision	Target Error Ratio
c-run3	15	19
c-run2	17	20
p-baseline	17	20

Table 1: Number of events meeting defined goals. The goals are defined to be met if the system's $PMiss < 4\%$ and $PFA < 50\%$.

3.3 Evaluation Results

Figure 3 shows the performance of our submitted runs among all submissions in term of actual NDC (Normalized Detection Cost) metric. The run c-run3 employs mainly the static visual, audio and semantic information in the videos. The run c-run2 employs the approach proposed in [12] and combines the motion and static features with late fusion. The actual NDC value is reduced from 0.57 to 0.47. This shows that motion information is useful in detecting the events defined by MED, especially for those events with intensive motion such as *bike trick*, *parkour*, *winning a race*, and *grooming animal*. Our proposed feature has the following advantages: i) It is not only motion, but also interactions between participants of an event which is important in describing the event. ii) It integrates two aspects of an event, i.e. static and motion information, in one single feature. Compared with other static or motion features, it completely describes an event. iii) Motion relativity eliminates the effect from camera motion, which is seldom addressed in most motion features. iv) The feature extraction is robust and not affected by complex background or environments. The keypoints are detected indepently and then combined in motion relativity. Once the keypoint detection is robust (which has been proven to be the case in many applications), so is the motion feature extraction.

The run p-baseline further employs the above mentioned feature selection methods to eliminate the less useful visual words and motion features. According to our experiments, the computation time is significantly reduced. At the same time, the performance is a little bit improved. This is because the original feature contains too much irrelevant information, which is useless, and even noisy sometimes in event classification. By selecting a rather clean set of features, larger weights are assigned to the most important features and the performance is thus improved.

Table 1 shows the number of events for which our systems meet the defined goals. By employing motion features, the two runs c-run2 and p-baseline meet the goals for more events with both threshold selection methods. In this work, thresholds are selected on development set by minimizing the NDC values. The resulting thresholds are basically good (although not the best) for test set.

Due to the lack of detailed results from MED evaluation on test set, Figure 4 shows the per-event performance of our systems on the development set. As can be observed in Figure 4, by employing motion features, consistent improvement is achieved for different events. For those events with intensive motion and interactions between event participants such as *Grooming Animal*, *Parade*, *Unstuck Vehicle*, *Parkour*, and *Cleaning Appliance*, the performance improvement is more significant. By feature selection, the run p-baseline outperforms c-run2 for most events. Although slight decline on performances can be found for few events, the approach basically works well at improving computation efficiency and also the overall accuracy of event detection.

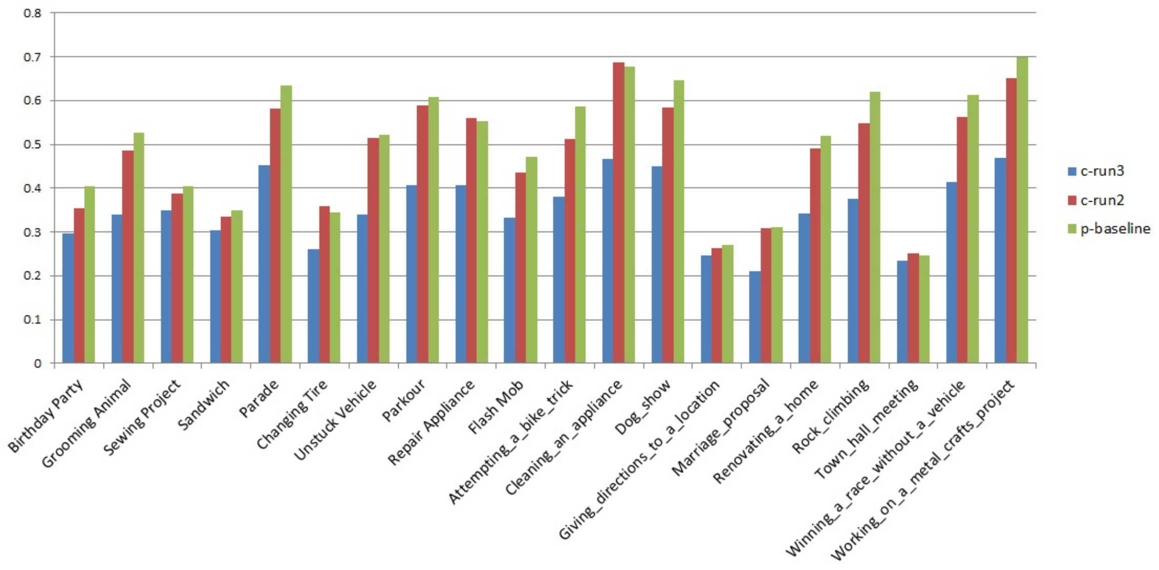

Figure 4: The performances of our submitted runs on the development set.

4 Conclusion

In this year's TRECIVD evaluation, we focus on validating our approaches for SIN and MED tasks. For SIN, we demonstrate the performance improvement achieved by the Weighted Hamming Embedding kernel. Compared with most current approaches, Hamming Embedding provides a more precise distance measure between different images for concept detection. By further weighting the informativeness of visual words, the more important features can be selected to boost the performance. This approach can also be extended to other classifer training stages as a method of feature selection. For MED, our motion features are demonstrated to be effective in describing and detecting video events. Comparisons with other motions features are to be done in future work.

Acknowledgement

The work described in this paper is partially sponsored by the National Natural Science Foundation of China (No. 61103127), and Shanghai Pujiang Program (No. 12PJ1402700).

References

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

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

[3] 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.

[4] 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.

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

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

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

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

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

[10] F. Wang and B. Merialdo, "Eurecom at TRECVID 2009: High-Level Feature Extraction", *TRECVID Workshop*, 2009.

[11] H. Jegou, M. Douze and C. Schmid, "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search", *European Conf. on Computer Vision*, 2008.

[12] F. Wang, Y. G. Jiang, and C. W. Ngo, "Video Event Detection Using Motion Relativity and Visual Relatedness", *ACM Multimedia*, 2008.

[13] F. Wang and B. Merialdo, "Weighting Informativeness of Bag-of-Visual-Words by Kernel Optimization for Video Concept Detection", *Int. Workshop on Very-Large-Scale Multimedia Corpus, Mining and Retrieval*, 2010.

[14] Y. G. Jiang, C. W. Ngo, and J. Yang, "Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval", *Conf. on Image and Video Retrieval*, 2007.

[15] J. Gemert, C. Veenman, A. Smeulders, and J. Geusebroek, "Visual Word Ambiguity", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2010.

[16] Y. G. Jiang and C. W. Ngo, "Bag-of-Visual-Words Expansion Using Visual Relatedness for Video Indexing", *ACM SIGIR*, 2008.

[17] P. Natarajan, *et. al*, "BBN VISER TRECVID 2011 Multimedia Event Detection System ", *TRECVID Workshop,* 2011.

[18] Y. G. Jiang, X. Zeng, G. Ye, D. Ellis, and S. F. Chang, "Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching", *TRECVID Workshop,* 2010.

[19] D. Xu and Shih-Fu Chang, "Video Event Recognition using Kernel Methods with Multi-Level Temporal Alignment", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no.11, pp. 1985-1997, 2008.

[20] Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582-1596, 2010.

[21] Usman Niaz, Miriam Redi, Claudiu Tanase, Bernard Merialdo, "EURECOM at TrecVid 2012: The Light Semantic Indexing Task", *TRECVID Workshop*, 2012.

[22] N. Cristianini, J. Kandola, A. Elisseeff, and J. S-Taylor, "On Kernel Target Alignment", *Advances in Neural Information Processing Systems*, vol. 14, pp. 367-373, 2002.