MULTIMEDIA EVENT DETECTION (MED) EVALUATION TASK

Matthias Dantone, Kenneth Sullivan, Jelena Tešić

Mayachitra Inc. Santa Barbara, CA, USA

ABSTRACT

Mayachitra Inc. team submitted runs for the TRECVID 2010 Multimedia Event Detection Pilot (MED) task evaluation. In this paper, we describe the preliminary set of results. The focus of this experiment for the Mayachitra Inc. team was to implement an end-to-end pilot system for multimedia event detection that (i) processes video,extracts and stores state-of-art video descriptors (ii) learns complex event models, and (iii) evaluates them on the test set in an efficient and effective manner. In this preliminary report, we summarize our findings on the performance of one of the important system components: the state-of-art activity detection approach. We have submitted two runs to NIST:

- **c_raw_1**: max-type fusion of the scores from binary detectors trained on the subset of visual words.
- **p_base_1**: weighted fusion of the individual scores from activity detector.

and evaluated additional run:

- **c_sel_1**: cross-validation fusion of the activity detectors trained on the expanded set

The performance of the runs varied significantly based on the training selection, and diversifying training set improves the detection scores. Overall, the activity recognition component has definitely showed potential in the overall event detection system for user-generated video collections. We will present a detailed analysis in the final notebook paper.

1. ACTION DESCRIPTORS

Following the explosion of user-created video content, and lack of tools to efficiently index and retrieve them, the research community has made significant progress in advancing the use of static descriptors (i.e. visual descriptors extracted from video keyframes) to detect objects and scenes in automatic annotation pipeline, and to connect them to the events they describe [1, 2]. To describe a complete event, descriptors need to capture scene, objects, and their relations present, and the actual activity/action. The research effort of incorporating the activity recognition analysis in a scalable video analysis systems is still in its infancy.

Lately, the computer vision community reported favorable results in action recognition domain as it extended traditional object recognition approaches to the spatio-temporal domain of video dataset [3, 4]. The actions are captured as spatio-temporal patterns in the local descriptor space. To effectively capture the actions in the user-generated video content, such as YouTube video dataset, we must consider the following:

- The size of video archive is overwhelming.
- User-created video content is widely diverse in content capture (camera settings), content presentation (event flow), and content editing.
- Actions that need to be detected vary in scale of details that need to be captured.

This boils down to the following demands on the selection of the state-of-art spatio– temporal descriptor: (i) the descriptor extraction needs to be efficient (ii) the features extracted need to be time and scale invariant, (iii) the extracted features needs to capture rich semantics of action events in video archives. For the TRECVID MED pilot task, we use the dense, scale-invariant, spatio-temporal Hes-STIP detector of Willems et al. [5]. This detector responds to spatio-temporal blobs within a video, based on an approximation of the determinant of the Hessian. These features are scale-invariant (both in temporal and spatial domain), and relatively dense comparing with other spatio-temporal features.

1.1. Spatio-temporal interest point detection

The spatio-temporal scale space L is defined by a spatio-temporal signal f convolving with a Gaussian kernel $g(\cdot; o_2, r_2)$, where o represents the spatial and r the temporal scale.

$$L(\cdot; o_2, r_2) = g(\cdot; o_2, r_2) * f(\cdot)$$

Willems et al. [5] used the Hession Matrix for the point detection task. The Hession Matrix H is defined as the square matrix of all second-order partial derivatives of L.

$$H = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{pmatrix}$$

The Gaussian second-order derivatives in the spatio- temporal space (D_{xx}, D_{yy}, D_{tt}, D_{xy}, D_{tx} and D_{ty}) can be approximate using box-filters [6]. All six derivatives can be computed by rotated version of only two different types of box filters. The box filters can be calculated efficiently using an integral representation of the video, [7]. The determinant of the matrix H defines the strength of a point of interest at certain scale.

1.2. SURF3D Descriptor

The descriptor used by Willems et al. is an extension of the 2D SURF image descriptor [6]. To describe the interest point a rectangular volume with the dimension $so \times so \times sr$ must be defined, where r represents the time scale, o the spatial scale and s is a magnification factor. The descriptor volume is divided into $M \times M \times N$ subregion. Within each of these sub volumes 3 axis-aligned Box-Filters d_x, d_y, d_t are calculated at uniform sample points. Every subregions is represented by the vector $v = (\sum d_x, \sum d_y, \sum d_y)$. The resulting descriptor is invariant to spatial rotation if the dominant orientation has been taken into account and he is invariant to spatial and temporal scale if the used Box-Filters have had the size $o \times o \times r$. We use this dense, scale-invariant, spatio-temporal HesSTIP detector and SURF3D descriptor in our activity detection pipeline.

2. ACTIVITY RECOGNITION

An event for MED 2010 is "an activity-centered happening that involves people engaged in process-driven actions with other people and/or objects at a specific place and time". In this preliminary report, we present the activity recognition component of our system.

2.1. Training

We used the activity descriptors extracted from the development set, as described in Section2, to build a visual vocabulary of size $N = 200$, as described in [8]. Every video is then represented by its histogram over the visual words vocabulary.

To train the activity detectors, we have used the libSVM package [9] to train the activity recognition models on top of the visual word descriptors, and learn nonlinear decision boundaries in activity descriptor space. Development data set consists of 1746 videos: 150 labeled videos – 50 instances of each of the three MED '10 events ("making a cake", "batting a run", and "assembling a shelter") and the rest of the video clips (1596) do not include any of the three events of interest. For each of the three events, there are 50 videos that contain that event, and 1696 videos that do not. We adopt the approach in [10] and build a set of base SVM classifiers for each class. We use the class examples as positive data points for each primitive classifier, and we use a different set of video example points for each of these models, thus leveraging the underlying semantics to expand on the diversity of the negative examples for each of the base classifiers.

To avoid the over-fitting or unbalanced learning scenarios, we subsample the negative data points. We have used two different settings: in the first scenario (**raw** and **base** run), we have trained the base classifiers using the class examples as positives and other class labels and small subsample of unlabeled data as negatives; in the second scenario (**sel** run), we have extended the negative sampling, and combined the labeled and unlabeled data to create the negative set. For system performance we only selected to train and evaluate a handful of base SVM classifiers (3), rather than utilizing the whole dataset and evaluating a higher number of base SVM classifiers (25).

Since production features vary significantly even for the development video set, we have employed several techniques to minimize the sensitivity of the modeling to production factors. In the training process, we selected the optimal set of SVM parameters using grid search strategy. The optimal learning parameters are selected based on the performance measure on the same 5-fold cross validation on training data.

3. EVALUATION

The goal of this exercise was to evaluate state-of-art video extraction approaches, and assess its potential contribution to an end-to-end large scale video analysis system in terms of performance scalability and accuracy. The approach was evaluated on TRECVID MED 2010 collection. The development dataset consists of 1746 videos, with duration close to 56 hours, and the evaluation dataset consists of 1742 clips with duration close to 59 hours. Note that frame size, frame rate and length of the videos varies. To increase the efficiency of our system, we have rescaled all videos in development and evaluation set to frame size of *160x120* pixel. The scores from each base detectors are fused using (a) max score and (b) weighted average, where weights for the base detectors were learned through cross-validation on the development set.

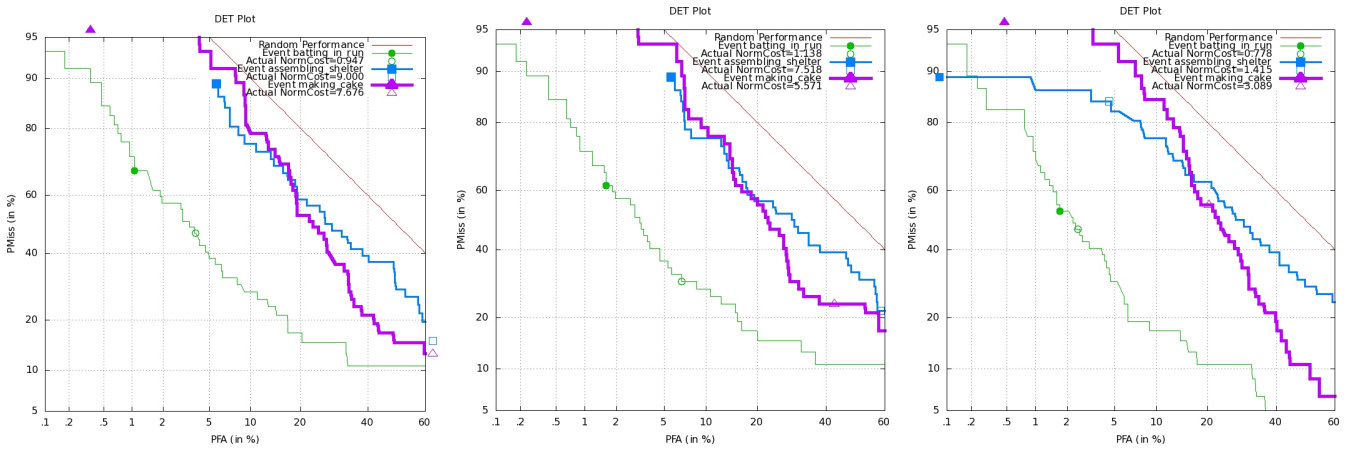
The evaluated runs are defined as follows:

- **raw**: max-type fusion of the binary detector scores
- **base**: cross-validation fusion of the base detections
- **sel**: cross-validation fusion of the improved detections

The evaluation summary of the three runs is outlined in Table 1. Figure 1 show the performance evaluation graphs of the c_raw_1, p_base_1, and c_sel_1 runs, respectively.

	class	Act. PFA	Act. PMiss	NormCost
raw	run	0.0384	0.4681	0.9472
	cake	0.6045	0.1277	7.6762
	shelter	0.7086	0.1522	9.0002
base	run	0.0673	0.2979	1.1382
	cake	0.4274	0.2340	5.5711
	shelter	0.5847	0.2174	7.5183
sel	run	0.0248	0.4681	**0.7777**
	cake	0.2031	0.5532	**3.0890**
	shelter	0.0454	0.8478	**1.4151**

Table 1. Summary of the performance results for the three submitted runs



(a) raw run (b) base run (c) sel run

Fig. 1. Detection error trade-off graphs of the activity descriptor runs

In the **raw** and **base** runs, the selection of negative data points was evenly distributed among other two labeled events, and randomly sampled remainder of the development set. This resulted in three base SVM detectors for each run. The descriptor scores were fused using max criteria for **raw** run and learned weights for the **base** run. Note that the cross-validation fusion improves over the max fusion approach, as shown in Table 1. The detection pipeline shows to be more sensitive to the selection of negatives for the base SVM models, than to the fusion technique. In the **sel** run, each base detector negative samples contained the equal number of unlabeled and labeled data from the development set. This resulted in better base modeling results, and final sel run is superior to the base and raw runs.

4. CONCLUSION

Mayachitra Inc. team participated in the TRECVID 2010 Multimedia Event Detection Pilot task. In this paper, we present the preliminary results and experiments conducted using our pilot system implementation. The goal of the exercise was to evaluate how much activity descriptors can contribute to an overall scalable multimedia search system. More details on the overall system performance and related analysis will be provided in the final notebook paper.

5. REFERENCES

[1] M. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia Magazine*, vol. 13, no. 3, 2006.

[2] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, pp. 151–174. 2009.

[3] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] Saad Ali and Mubarak Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288–303, 2010.

[5] Geert Willems, Tinne Tuytelaars, and Luc J. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proceedings of ECCV, 10th European Conference on Computer Vision*, 2008, pp. 650–663.

[6] H. Bay, Tinne Tuytelaars, and Luc J. Van Gool, "Surf: Speeded up robust features," in *Proceedings of ECCV, 9th European Conference on Computer Vision*, 2006, pp. 404–417.

[7] Yan Ke, Rahul Sukthankar, and Martial Hebert, "Efficient visual event detection using volumetric features," in *Proceedings of 10th IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 166–173.

[8] K. Sullivan, S. Chandrasekaran, K. Solanki, B. S. Manjunath, J. Nayak, and L. Bertelli, "Hierarchical scene understanding exploiting automatically derived contextual data," in *Proceedings of SPIE Defense, Security, and Sensing*, 2010.

[9] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines," 2001, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[10] J. Tešić, A. Natsev, and J. R. Smith, "Cluster-based data modeling for semantic video search," in *ACM International Conference on Image and Video Retrieval (ACM CIVR)*, 2007.