MMM-TJU at TRECVID 2010

An-An Liu, Zan Gao, Yuting Su, Zhaoxuan Yang
Multimedia Mining Group, School of Electronic Information Engineering,
Tianjin University, Tianjin 300072 P.R. China

Abstract

Surveillance Event Detection

Semantic event detection in the huge amount of surveillance video in both retrospective and real-time styles is essential to a variety of higher-level applications in the public security. In TRECVID 2010, to overcome the limitations of the traditional human action analysis method with human detection/tracking and domain knowledge, we evaluate the general framework for multiple human behaviors modeling with the philosophy of ***bag of spatiotemporal feature (BoSTF)***. The brief introduction to each run is shown in Table 1.

Table 1 SED RUNs ActDCR Description

SED_Runs	ActDCR	Descriptions
TJUMM_1	6.6931	MoSIFT, SVM with χ^2 kernel, 0.50 (threshold)
TJUMM_2	5.2067	MoSIFT, SVM with χ^2 kernel, 0.60 (threshold)
TJUMM_3	4.4773	MoSIFT, SVM with χ^2 kernel, 0.65 (threshold)
TJUMM_4	3.7913	MoSIFT, SVM with χ^2 kernel, 0.70 (threshold)
TJUMM_5	3.1070	MoSIFT, SVM with χ^2 kernel, 0.75 (threshold)
TJUMM_6	2.4753	MoSIFT, SVM with χ^2 kernel, 0.80 (threshold)
TJUMM_7	1.9196	MoSIFT, SVM with χ^2 kernel, 0.85 (threshold)
TJUMM_8	1.4527	MoSIFT, SVM with χ^2 kernel, 0.90 (threshold)

Semantic Indexing

Semantics indexing is extremely helpful for automatic semantic discovery and annotation. In TRECVID 2010, we mainly evaluate three kinds of features, two global features (grid-based color moments and texture feature) , and one local feature (Scale-Invariant Feature Transform, SIFT) for semantic modeling. The cascade Support Vector Machine (SVM) is implemented for each concept model leaning in two ways. First, for each concept three kinds of classifiers are learned with individuals. Second, the decision of three experts above are fused with average fusion algorithm to take advantage of the superiority of individuals. Therefore, we obtained four runs for evaluation of the High Level Feature Extraction test in TRECVID 2010. The brief introduction to each run is shown in Table 2.

Table 2 SIN RUNs infMAP Description

SIN_Runs	InfMAP	Descriptions
L_A_MMM-TJU1_1	0.0156	GCM Feature and Cascade SVM
L_A_MMM-TJU2_2	0.0052	Texture Feature and Cascade SVM
L_A_MMM-TJU3_3	0.0238	SIFT Feature and Cascade SVM
L_A_MMM-TJU4_4	0.0267	Fusing All Results

TRECVID 2010 Surveillance Event Detection by MMM-TJU*

An-An Liu[1†], *Zan Gao*[1,2], *Yuting Su*[1], *Zhaoxuan Yang*[1]

[1] *School of Electronic Information Engineering, Tianjin University, Tianjin* 300072 *P.R. China*
[2] *School of Information and Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China*

Abstract

Efficient and robust human behavior detection in vast amounts surveillance video in real time is the fundamental technology for a variety of higher-level applications of public security. Thus, TRECVID supplies the chance to compare different schemes of surveillance event detection from with public dataset and rules. The goal of the evaluation track is to support the development of technologies to detect people-engaged visual events in a large collection of streaming video data. TRECVID 2010 Surveillance Event Detection evaluation is operated with 150 hours of multi-camera airport surveillance domain data collected by the Home Office Scientific Development Branch (HOSDB) and the ground-truth by the University of Pennsylvania Linguistic Data Consortium. The development data of the evaluation consists of the 2008 Event Detection training and test sets. The evaluation set is the UK Home Office Scientific Development Branch's (HOSDB) i-LIDS MCTTR dataset. It is the same evaluation corpus as used for the 2009 evaluation. The events of interest include seven items, CellToEar, Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns, and Pointing. MMM-TJU implemented the general framework to model human behavior with the philosophy of ***bag of spatiotemporal feature (BoSTF)*** for individual event.

1. System Framework

Our team utilized the general framework to model human behavior with the philosophy of ***bag of spatiotemporal feature (BoSTF)*** for individual event as shown in Figure 1. For each temporal sliding window in the video sequence, the MoSIFT interest points are detected and formulated. Then the extracted MoSIFT points are clustered into visual keywords and SVM classifier is used for semantic event modeling. In the evaluation the video sequences are tested in the same way and the temporal adjacent events from different neighbor sliding windows are fused as identitical one. With these post-processing results the final decision will be given for the entire sequence. Figure 2 shows our MoSIFT features in a Gatwick video key frame. It shows that MoSIFT features are able to clearly focus on areas with human activity.

* This work was supported in part by the Tianjin Research Program of Application Foundation and Advanced Technology (10JCYBJC25500), Doctoral Fund of Ministry of Education of China (20090032110028), Innovation Foundation of Tianjin University.
† Corresponding author

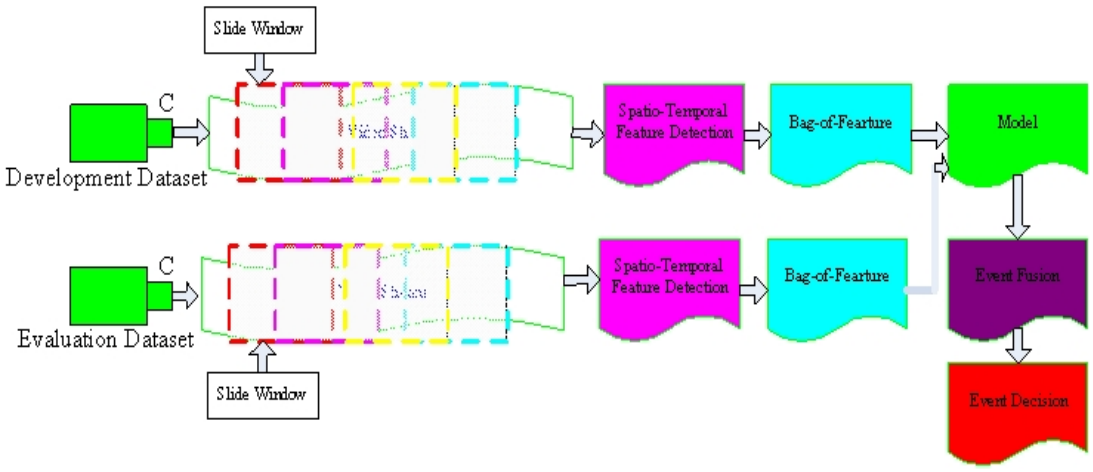


Figure 1 the framework of our surveillance event detection

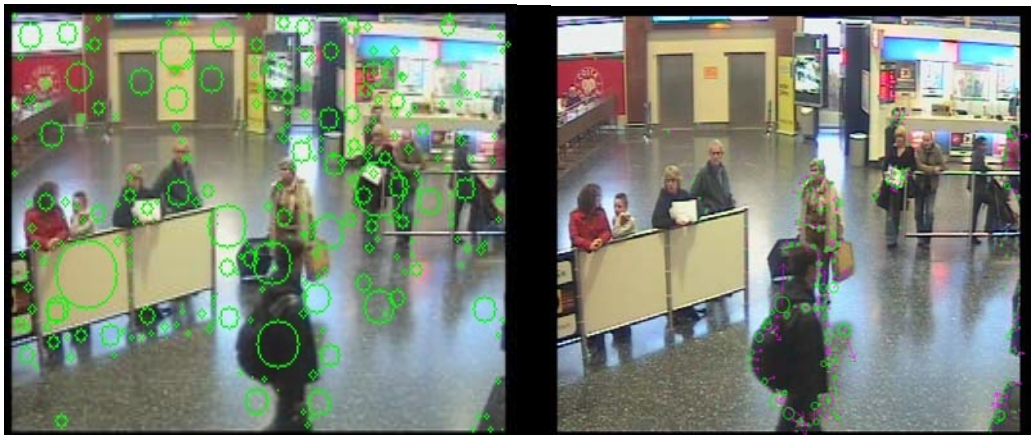
2. The Spatiotemporal Feature

To avoid the dependence on scenarios, appearance and so on, we utilize interest points for video feature description. Interest point extraction can transfer the video data from a large volume of pixels to a sparse but descriptive set of features. Ideally, an interest point detector should densely sample those portions of the video where events occur while avoiding regions of low activity. Therefore, we use MOSIFT, the popular 3D point feature for this task. The philosophy behind our MoSIFT interest point detector is to treat appearance and motion separately, and to explicitly identify spatially-distinctive regions in a frame that exhibit sufficient motion at a variety of spatial scales [1].

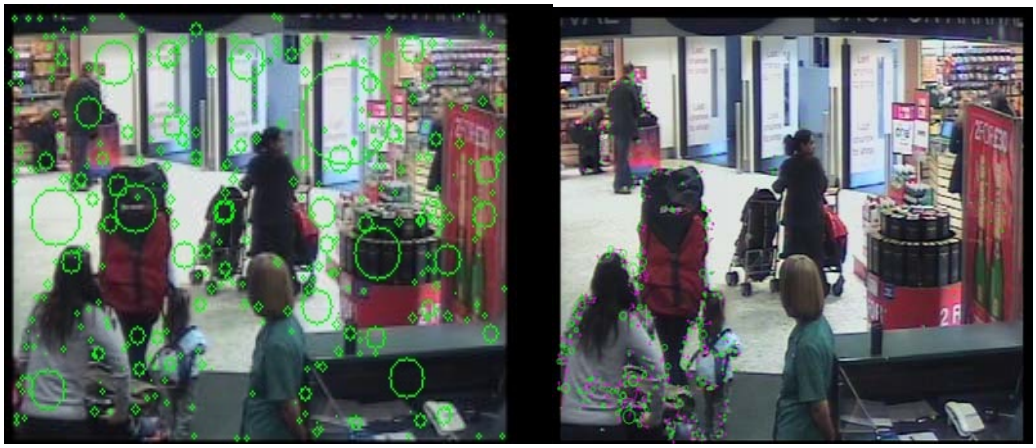
Like other SIFT-style keypoint detectors, MoSIFT finds interest points at multiple spatial scales. Two major computations are employed. First, SIFT interest point detection on the first frame to identify candidate features. Second, optical flow computation between the two frames, at a scale appropriate to the candidate feature, to eliminate those candidates that are not in motion. Candidate interest points are determined using SIFT point extraction method [2] on the first frame of the pair by the local extrema (minima/maxima) of the DoG images. Then, the algorithm scans through each octave and interval in the DoG pyramid and extracts all of the possible interest points at each scale. At last, the candidate points are checked against the optical flow pyramid and some of them are selected as MoSIFT interest points only if they contain sufficient motion in the optical flow pyramid at the appropriate scale. The samples are shown in Figture 3.

"***Bag of Words (BoW)***" is a popular method for representing documents in Natural Language Processing. In recent years, lots of researchers in computer vision are engaged in utilizing this method for object recognition and classification [3]. To represent an image using ***BoW*** model, an image/video can be treated as a document. We can utilize the key points and its description detected in Section A as "visual word". Then, the main work of this step is to generate the codebook and to convert vector represented interest points to vocabulary for spatiotemporal feature construction. A codeword can be

considered as a representative of several similar patches. We performed K-means algorithm [4] over all the vectors. Vocabularies are then defined as the centers of the learned clusters. The number of the clusters is the codebook size. Thus, each detected point in an image/video is mapped to a certain codeword through the clustering process and the spatiotemporal feature of the image/video can be represented by the histogram of the vocabulary.



(a) Sample 1



(b) Sample 2

Figure 2. Interest points detected with SIFT (left) and MoSIFT (right). Green circles denote interest points at different scales while magenta arrows illustrate optical flow. Note that MoSIFT identifies distinctive regions that exhibit significant motion, which corresponds well to human activity while SIFT fires strongly on the cluttered background.

3. Experiments and Discussion

In our experiment, the size of the slide window is set with experienced value 30 frames per second and the temporal step is the experienced value, 10 frames per second. The vocabulary size is 2000 depending on our pervious experiments. For each sliding window, all of the spatio-temporal interest

points in the window are projected into the vocabulary, and then spatiotemporal feature of the window can be represented by the histogram of the vocabulary.

In the training set, annotations are distributed to each window to mark it as positive or negative. This creates a highly unbalanced dataset (positive windows are much less frequent than negative windows). Thus, the development 2008 is used to trained, and evaluation 2008 is used to validate the parameter of SVM with χ^2 kernel [5,6,7]. We also aggregated consecutive positive predictions to achieve multi-resolution. The detection result is shown in the table 3. Figure 4 denotes the performance of the DET curve of TJUMM-8. From table 3, five of seven events are less than 1 in MinDCR, which is informally equivalent to random performance.

Table 1 RFA denotes Rates of False Alarms. PMiss denotes probability of missed event. DCR denotes Detection Cost Rate.

Analysis Report	#Ref	#Sys	#CorDet	#FA	#Miss	Act. RFA	Act. PMiss	Act. DCR	Min RFA	Min PMiss	Min DCR
CellToEar	194	789	8	781	186	51.222	0.959	1.215	0.066	1.000	1.000
Embrace	175	3672	86	3586	89	235.190	0.509	1.684	29.645	0.834	0.983
ObjectPut	621	885	15	870	606	57.060	0.976	1.261	0.066	1.000	1.000
PeopleMeet	449	3482	152	3330	297	218.400	0.661	1.753	2.230	0.969	0.980
PeopleSplitUp	187	1515	28	1487	159	97.526	0.850	1.338	2.755	0.984	0.998
PersonRuns	107	3217	49	3168	58	207.776	0.542	1.581	7.477	0.925	0.963
Pointing	1063	1441	107	1334	956	87.491	0.899	1.337	0.066	1.000	1.000

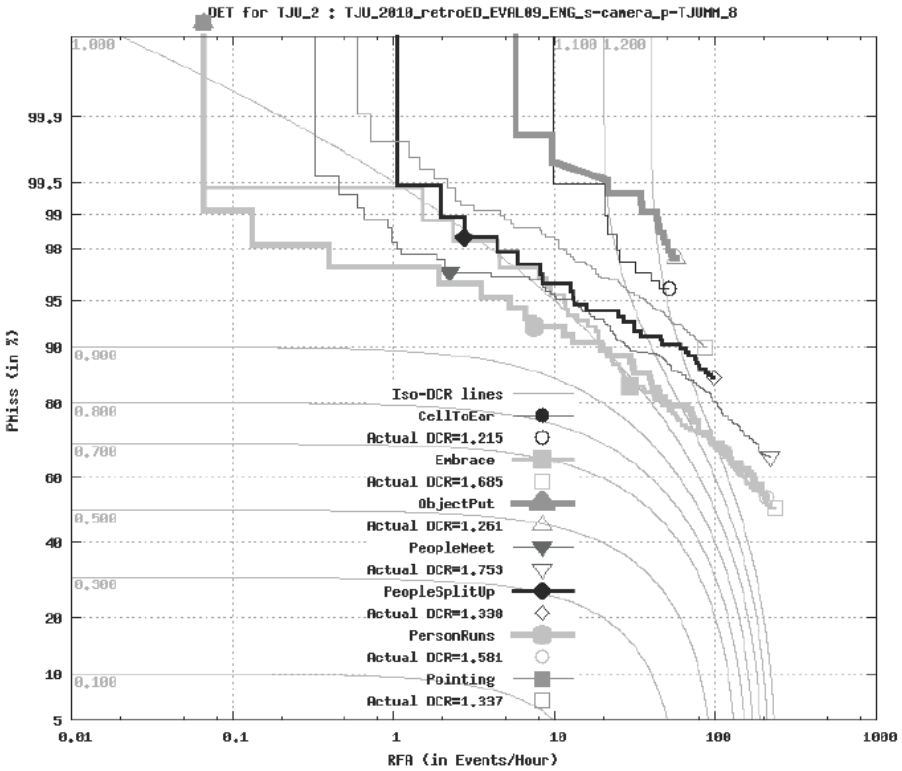


Figure 3 The DET curve of our TJUMM-8

4. Conclusions

In TRECVID 2010 we evaluate the general framework for multiple human behaviors modeling with the philosophy of ***BoSTF***. The experimental reasults show that facing the variant visual pattern of the same human action and the complicated visual pattern of multiple behavious this general framework would overcome the limitations of the tranditional rule-based human action detection method. However, to improve the detection accuracy in the future more disriminated features could be fomulated for better visual representation and temporal inference models could be tried to make good use of the potential temproal information.

Reference:

[1]. Ming-yu Chen and Alex Hauptmann. MoSIFT: Reocgnizing Human Actions in Surveillance Videos . CMU-CS-09-161, Carnegie Mellon University, 2009.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[3] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. IEEE Comp. Vis. Patt. Recog. 2005.

[4] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV, 43(1):29–44, June 2001.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[6] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. ML, 68(3):267–276, 2007.

[7] J. C. Platt. Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, editors, Advances in Large Margin Classifiers, pages 61–74. The MIT Press, 2000.

TRECVID 2010 Semantic Indexing by MMM-TJU*

Zan Gao [1,2] *, An-An Liu* [1‡] *, Yuting Su* [1] *, Zhaoxuan Yang* [1]

[1] *School of Electronic Information Engineering, Tianjin University, Tianjin* 300072 *P.R. China*
[2] *School of Information and Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China*

1. Abstract

In TRECVID 2010, we mainly evaluate three kinds of features, two global features (grid-based color moments and texture feature) , and one local feature (Scale-Invariant Feature Transform, SIFT) for semantic modeling. The cascade Support Vector Machine (SVM) is implemented for each concept model leaning in two ways. First, for each concept three kinds of classifiers are learned with individuals. Second, the decision of three experts above are fused with average fusion algorithm to take advantage of the superiority of individuals. Therefore, we obtained four runs for evaluation of the High Level Feature Extraction test in TRECVID 2010.

2. Low-level feature extraction

In semantic indexing of TRECVID 2010 approximately 8000 Internet Archive videos (50GB, 200 hours) with Creative Commons licenses in MPEG-4/H.264 lasting between 10 seconds and 3.5 minutes are released for algorithm development. Both training dataset and test dataset (IACC.1.A) are 200 hours drawn from the IACC.1 collection using videos with durations between 10 seconds and 3.5 minutes. As the dataset is from internet, it is impossible for us to represent their saliency content only with one kind of feature. Therefore in our system, three kinds of low level visual features, grid-based color moments, texture and SIFT feature, are extracted.

2.1 Grid-based color comments (GCM)

This color descriptor is most suitable for representing local (object or image region) features where a small number of colors are enough to characterize the color information in the region of interest. In addition, we choose color feature detectors according by MPEG-7, and select some color detectors which have been proved that they are effective. To generate the color moment feature, each image (key-frame) is divided into 5x5 grids, and each grid is described by the mean, standard deviation, and third root of the skewness of each color channel in the LUV color space. This results in a 225-dimension (5x5x3x3) color moment feature.

2.2 Texture feature

* This work was supported in part by the Tianjin Research Program of Application Foundation and Advanced Technology (10JCYBJC25500), Doctoral Fund of Ministry of Education of China (20090032110028), Innovation Foundation of Tianjin University.

‡ Corresponding author

The Gabor wavelets can be defined as follows [10]:

$$\psi_{\mu,\nu}(z) = \frac{\left\| k_{\mu,\nu} \right\|^2}{\sigma^2} e^{-\frac{\left\| k_{\mu,\nu} \right\|^2 \bullet \|z\|^2}{2\sigma^2}} [e^{ik_{\mu,\nu}z} - e^{-\frac{\sigma^2}{2}}] \quad (1)$$

where μ and ν define the orientation and scale of the Gabor kernels, $z = (x, y)$, $\|\bullet\|$ denotes the norm operator, and the wave vector $k_{\mu,\nu}$ is defined as follows:

$$k_{\mu,\nu} = k_\nu e^{i\Phi_\mu} \quad (2)$$

where $k_\nu = k_{\max} / f^\nu$, $\Phi_\mu = \pi\mu/8$, $k_{\max}$ is the maximum frequency, and f is the spacing factor between kernel in the frequency domain. In most cases one would use Gabor wavelets at five different scales, $\nu \in (0,1,....,4)$, and eight orientations, $\mu \in (0,1,.....7)$. However, in order to save time, we just let $\nu \in (0,2,4)$ and $\mu \in (0,1,...5)$, at the same time, we choose the following parameters: $\sigma = 2.5\pi$, $k_{\max} = \pi/2$ and $f = \sqrt{2}$.

Let $I(x, y)$ be the gray level distribution of an image, and define the convolution output of image I and a Gabor kernel $\Psi_{\mu,\nu}$ as follows:

$$O_{\mu,\nu}(z) = I(z) \otimes \Psi_{\mu,\nu}(z) \quad (3)$$

where $z = (x, y)$, and $\otimes$ denotes the convolution operator. After we get the $O_{\mu,\nu}(z)$, we compute it in 7*7 image grids. In each grid we use the mean and variance of twelve oriented energy filters aligned in 30-degree intervals.

2.3 SIFT feature

In [5, 6, 7, 8, 9] SIFT feature achieved very promising performance in semantic modeling. Therefore we select SIFT feature for evluation. The local feature of each image is computed from the local key points detected from the image. We use the key points using the DoG detector and depicted by Scale-invariant feature transform (SIFT) descriptors [1] which describes each key points by a 128-dimension vector. SIFT features are invariant to image scale and rotation, and are also robust to changes in illumination, noise, occlusion and minor changes in viewpoint. For each key frame, the number of extracted key points is different. Therefore, we try to use bag-of-words (***BoW***) to quantify SIFT feature to a fixed number vector feature of each key frame. We use K-means clustering to find the conceptual meaningful clusters and each cluster is treated as a visual word in ***BoW*** approach. All the visual words consist of a visual word vocabulary. Then key points in each key frame are assigned to clusters in the visual vocabulary which are their nearest neighbors. In the end, each key frame is presented by a visual word histogram feature. As for vocabulary size and weighting scheme, our previous work shows using a moderate visual word vocabulary size lead a better performance. Therefore we cluster the key points into 2000 clusters. And the soft-weight scheme is adopted in our experiments. For each key point in an image, we select N (N=4) nearest neighbor clusters for it. These N nearest neighbor clusters are then assigned weights with their inverse rank value. The final weight

of each cluster is the sum of inverse rank values calculated from all the keypoints in an image.

3. System Framework

The framework is shown in Figure.1. Because the quality of videos from Internet Archive video with Creative Commons Internet vary greatly, three representative kinds of features are extracted for complement. By this way, we hope these complement. Due to the unbalance of the positive and negative samples, the cascade SVM [2] is trained for concept model learning. For each feature and concept, 10-layers cascade SVM is trained. In addition, different kernels in the SVM are adopted. For grid-based color moments and texture feature, the RBF kernel is employed while for SIFT the SVM with χ^2 kernel is trained. Finally, the output probabilities of cascade SVMs are fused to take advantages of the superiority of individuals. In our experiments, the decisions of multiple SVM classifiers with probabilistic output [3, 4] are fused with average fusion algorithm.

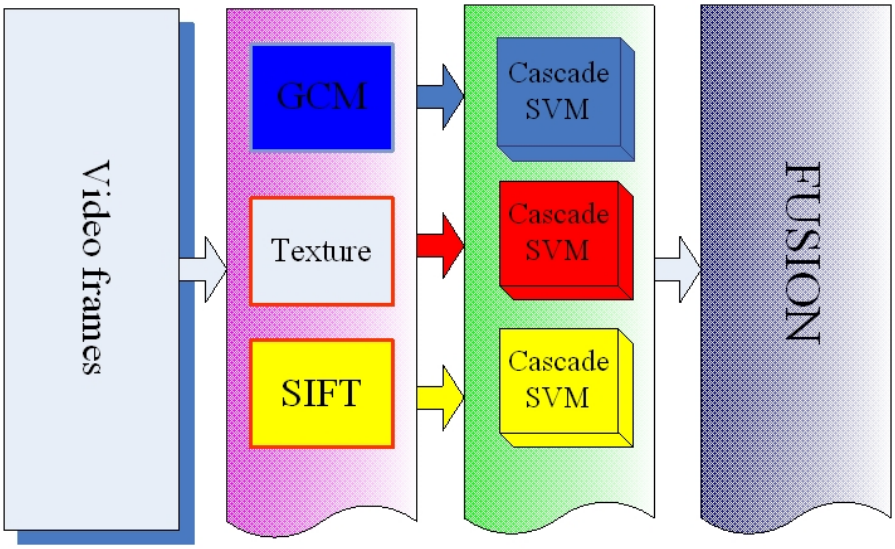


Figure.1 The framework of MMM-TJU

Table 1 SIN RUNs infMAP Description

SIN_Runs	InfMAP	Descriptions
L_A_MMM-TJU1_1	0.0156	GCM Feature and Cascade SVM
L_A_MMM-TJU2_2	0.0052	Texture Feature and Cascade SVM
L_A_MMM-TJU3_3	0.0238	SIFT Feature and Cascade SVM
L_A_MMM-TJU4_4	0.0267	Fusing All Results

4. Experiments and Discussions

In the TRECVID 2010 we submitted 4 runs by random selection for evaluation. The performance and description are shown in Table 1. From the Table 1, the performances of GCM,

Texture and SIFT are 0.0156, 0.0052 and 0.0238 respectively. The performance of SIFT feature is much better than that of GCM and Texture. The performance of Texture is the worst in all the features. Because the video content of different ones change drastically, texture feature is not discriminative enough for representation. When combining the output of cascade SVMs, the performance of combination is 0.0267, and the improvement of combination can reach 12.18% for SIFT feature. The average precision for each semantic concepts per run is shown in Figure.2, and MAPs for all runs are shown in Figuare.3. In Figure.2, the semantic index is defined as follows:
1-“Airplane_Flying”; 2-“Boat_Ship”; 3-“Bus”; 4-“Cityscape”; 5-“Classroom”; 6-“Demonstration_Or_Protest”; 7-“Hand”; 8-“Nighttime”; 9-“Singing”; 10-“Telephone”. In Figure.3 the horizontal axis is different runs index and the vertical axis is the mean average precision. From Figure.3 we can see that our top two MAP are above the average performance.

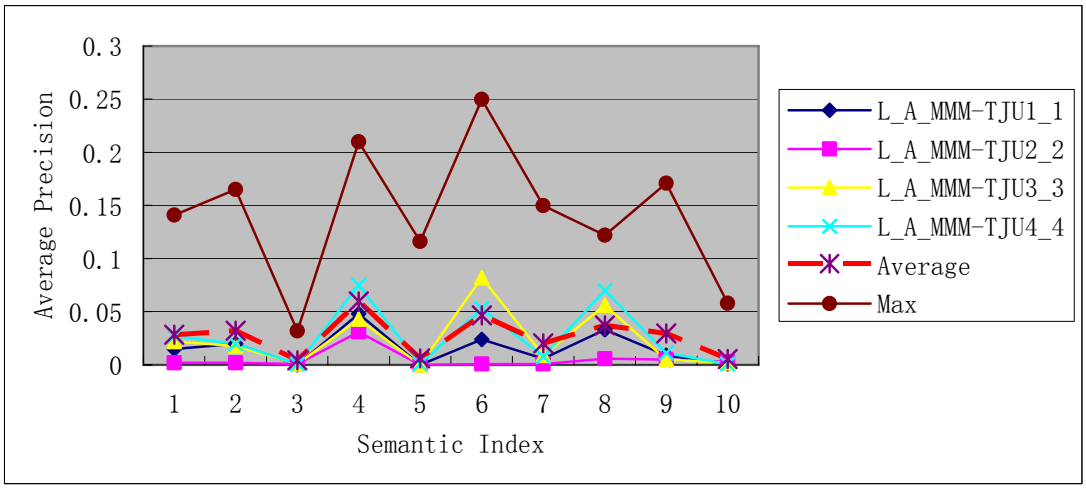


Figure.2 the Average Precision for each semantic per runs. Average and Max mean that we calculated the average and maximum of all the teams who submit the semantic indexing task.

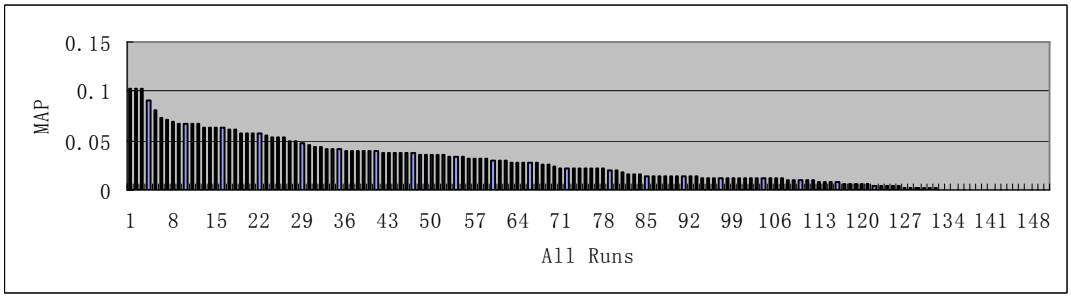


Figure.3 the performance of 4 submitted runs for Semantic Indexing. The red bars are from MMM-TJU.

5. Conclusions

In TRECVID 2010 we mainly evaluate three kinds of features, two global features (grid-based color moments and texture feature) , and one local feature (Scale-Invariant Feature Transform, SIFT) for semantic modeling. The cascade Support Vector Machine (SVM) is implemented for each concept model leaning. The experimental reasults show that features selection is crucial for semantic modeling and SIFT are much obviously more discriminative than color and texture feature. However, to improve the detection accuracy in the future more issues need advanced consideration: 1) more effective fusion strategies need test to reduce noise; 2) large-scale parallel computing and GPU would play an important role in semantic modeling to reduce computational intense.

Reference:

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[3] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. ML, 68(3):267–276, 2007.

[4] J. C. Platt. Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, editors, Advances in Large Margin Classifiers, pages 61–74. The MIT Press, 2000.

[5]. Sheng Tang, Yong-Dong Zhang, Jin-Tao Li et al. TRECVID 2007 High-Level Feature Extraction By MCG-ICT-CAS. In: Proceedings of TRECVID 2007 Workshop.

[6]. C.G.M. Snoek, I. Everts, J.C. van Gemert et al. The MediaMill TRECVID 2007 Semantic Video Search Engine. In: Proceedings of TRECVID 2007 Workshop.

[7]. James Philbin, Ondˇrej Chum, Josef Sivic et al. Oxford TRECVID 2007 – Notebook paper. In: Proceedings of TRECVID 2007 Workshop.

[8]. Yu-Gang Jiang, Chong-Wah Ngo, Jun Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007, 2007, p 494-501.

[9]. Van De Sande, Koen E. A. (ISLA, University of Amsterdam); Gevers, Theo; Snoek, Cees G. M. Evaluation of color descriptors for object and scene recognition, 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008, p 4587658

[10]. Liu, C., Wechsler, H. A Gabor Feature Classifier for face recognition, Proceedings of the IEEE International Conference on Computer Vision, v 2, 2001, p 270-275.