Top-Down Cues for Event Recognition

Li Li[1,2], Chunfeng Yuan[1], Weiming Hu[1], Bing Li[1]

[1] Institute of Automation, Chinese Academy of Sciences
[2] Radio, Film and Television Design and Research Institute

Abstract. How to fuse static and dynamic information is a key issue in event analysis. In this paper, we present a novel approach to combine appearance and motion information together through a top-down manner for event recognition in real videos. Unlike the conventional bottom-up way, attention can be focused volitionally on top-down signals derived from task demands. A video is represented by a collection of spatio-temporal features, called video words by quantizing the extracted spatio-temporal interest points (STIPs) from the video. We propose two approaches to build class specific visual or motion histograms for the corresponding features. One is using the probability of a class given a visual or motion word. High probability means more attention should be paid to this word. Moreover, in order to incorporate the negative information for each word, we propose to utilize the mutual information between each word and event label. High mutual information means high relevance between this word and the class label. Both methods not only can characterize two aspects of an event, but also can select the relevant words, which are all discriminative to the corresponding event. Experimental results on the TRECVID 2005 and the HOHA video corpus demonstrate that the mean average precision has been improved by using the proposed method.

1 Introduction

Event recognition is a key task in automatic video analysis, such as semantic summarization, annotation and retrieval. Since 2001, the National Institute of Standards and Technology (NIST) has started benchmarking content-based-video retrieval technologies, known as TRECVID [1], in which event detection is one of the evaluation tasks. NIST provides a benchmark of annotated video corpus for detecting a set of predefined concepts. Although a lot of efforts have been made for video based event recognition [2–4] and some preliminary results have been achieved during the past several years, the problem is still far away from being solved. This is mainly due to the within-event variations caused by many factors, such as unconstrained motions, cluttered backgrounds, occlusions, environmental illuminations and objects' geometric variances.

Recently, many researchers showed their interests in an approach that considers each video sequence as a collection of spatio-temporal interest points (STIPs). Laptev *et al.* [5] first incorporated the temporal constraint to a Harris interesting point detector to detect local 3D interesting points in the space-time dimension.

Dollar *et al.*[6] improved the 3D Harris detector and applied Gabor filtering to the spatial and temporal domain to detect interest points. In this method, a video can be modeled by the Bag of Words(BoW) [7]model, which has ability to handle variability in viewpoints, illumination and scales. This influential model represents each video as a collection of independent codewords in a pre-defined codebook generated from the training data.

In a video clip, an event usually has two important attributes: *what* and *how*. The *what* attribute usually refers to the appearance information obtained from static images. SIFT features [8] have been proved to be good candidates for the representation of image static information. Similar to SIFT features, Dalal and Triggs [9] proposed Histogram of Oriented Gradient (HOG) descriptors to handle pedestrian detection in static images. Recently, Scovanner *et al.* [10] proposed 3D SIFT features by applying sub-histograms to encode local temporal and spatial information. On the other hand, the *how* attribute refers to an event's dynamic information usually the object's motion. Motion feature has always been considered as an important cue to characterize an event. For instance, in [11], the event is modeled by volumetric features derived from optical flow in a video sequence. Zhang [12] extracted motion templates (motion images and motion context) using very simple processing. Histogram of oriented optical flow (HOOF) [13]was used to recognize human actions by classifying HOOF time series. Although the above existing approaches partially solved the even recognition in different aspects, how to effectively combine both *what* and *how* attributes is still an open problem for event recognition. To address this issue, in [2], a set of methods with motion and bag-of-visual-words combination were proposed to exploit the relativeness of the motion information and the relatedness of the visual information. Dalal *et al.* [9] combined motion and HOG appearance to achieve more robust descriptor. Efros *et al.* [14] employed appearance and flow features in an exemplar based detector for long shots of sports players, though quantitative performance results were not given.

The conventional approach representing a video usually adopts a bottom-up paradigm. In this work, we choose a top-down human visual system [15] instead to combine visual and motion cues for event recognition. In this top-down human visual system, only a subset of interesting information will be focused while the rest will be demoted. In other words, not all spatio-temporal features make the same contribution for different types of events. Some features may be useless for a particular class of events. If more weights given to the features highly relevant to the event recognition, the performance could be improved. Therefore, we propose two approaches in order to weight features: (1) Firstly, the probability of each word w.r.t the event classes is computed. Then a class-specific histogram is constructed so that the STIPs with the higher probability of the corresponding words under the category should be emphasized more; (2)As an alternative to the probability, the mutual information (MI) of each word w.r.t the event classes is computed. Mutual information is a nonparametric, model-free method for scoring a set of features. It can be used to spot all features relevant to the classification, and to identify groups of features that allow building a

valid classification model. Recently, MI has been proved to be effective way for the computation of visual recognition tasks. Liu and Shah [16] utilized the Maximization of Mutual Information (MMI) to automatically discover the optimal number of video word clusters. Yuan *et al.* [17] represented the video sequence as a bag of spatio-temporal invariant points (STIPs), where the MI between each STIP and a specific class was evaluated. In this method, action categorization is based on the mutual information maximization. Due to the independence assumption of STIPs, this model ignored the dependency among features. Different from [17], in this paper, we calculate the MI between each word rather than each STIP and specific class. This MI considers not only the positive effect of each word, but also the negative tone. By incorporating the negative training information, our model gains better discriminative power.

The rest of the paper is organized as follows. Section 2 presents the overall architecture of the proposed method. Section 3 describes the proposed video representation method and the class specific histogram in details. Section 4 provides experiment results. Finally, conclusions are given in Section 5.

2 Overall Architecture

This paper focuses on developing effective techniques for the combination of visual and motion features. Although more complicated appearance features such as 3D SIFT could be used instead and may achieve even better results, we adopt a relatively simple features, HOG (Histogram of Oriented Gradient) as visual descriptors and HOF (Histogram of Optical Flow) as motion descriptors in order to validate the proposed combination methods.

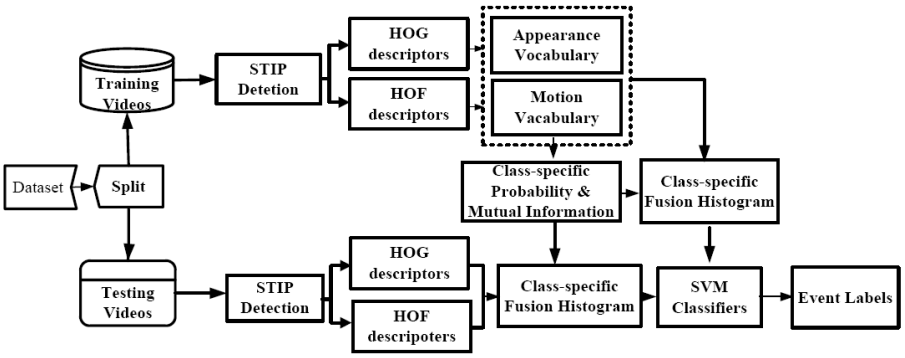

Fig. 1. The flowchart of the proposed approach for event recognition

Fig. 1 gives an overview of the framework. First of all, we employ Laptev *et al.* [5]'s method to detect spatio-temporal interest points (STIPs) in a video clip. After that, we utilize HOG as appearance descriptor and HOF as motion

descriptor. Subsequently, the visual and motion codebook are generated respectively by grouping the detected STIP features using the k-means algorithm. The center of each resulting cluster is defined as a video word. Subsequently, in order to combine visual and motion features, we build appearance based motion histogram and motion based appearance histogram in a top-down manner. For the learning process we use a method similar to [15]. We start with a set of training videos, in which all of the positive training sets have been manually marked. For example, HOG can be used as the descriptor cue and HOF can function as attention cue. Based on the probability and mutual information of the class for the given word, a class specific histogram is constructed. In this way, each video clip is eventually represented as an attention histogram in the framework of BoW. In the testing phase, the test video is also firstly represented as the attention histogram of BOW and then classified according to histogram matching between the test video and training videos. Finally, the test video is classified according to a SVM classifier with Histogram Intersection kernel.

In the next section, our descriptor and the attention histograms are described in details.

3 Top-Down Attention Histogram for Event Representation

3.1 STIPs: Video Representation

Bag-of-Words (BoW) [7, 18] model has been proved to be a powerful tool for various image analysis tasks. The visual vocabulary provides a mid-level representation which helps to bridge the semantic gap between the low-level features and the high-level concepts. We represent a video as a bag of spatio-temporal features $\{d_i\}$. Once the visual and motion codebook are generated, we represent a video clip by $\nu = \{w^k\}$ and $k \in \{a, m\}$ for the two cues appearance and motion codebook respectively. $W \in \omega = \{w_1^k, w_2^k, ..., w_n^k\}$ represents a set of video words. We denote by $\mathrm{T}^{c+} = \{\nu_i\}$ the positive training samples of class c. Symmetrically, T^{c-} is the negative training dataset of class c, and $\mathrm{T} = \mathrm{T}^{c+} \cup \mathrm{T}^{c-}$. For a standard single-cue BoW, videos are represented by the statistical distribution of BoW:

$$n(w^k|\nu) = \sum_{j=1}^{\|\nu\|} \delta(w_{d_j}^k, w^k) \tag{1}$$

where $\|\nu\|$ denotes the total number of STIPs in video ν, $\delta(\cdot)$ is the indication function, $w_{d_j}^k$ is the word index of the corresponding STIP d_j. Conventionally fusion methods of the two cues are called early fusion and late fusion respectively, while early fusion involves creating one joint appearance-motion vocabulary, and late fusion concatenates both histogram representation of both appearance and motion, obtained independently.

3.2 Top-Down Attention Histogram

Inspired by the recent work [15], in which a top-down color attention mechanism combines the advantages of early and late feature fusion together, we resort to the top-down human visual attention mechanism to recognize a specific event category $c \in \mathbf{C}$. Evidence from human vision indicates that high-level, class-based criteria play a crucial role in recognizing objects [19]. The computation of the video representation is done according to

$$n(w^a|\nu, \mathbf{C} = c) = \sum_{j=1}^{\|\nu\|} \pi(w^m_{d_j}, \mathbf{C} = c)\delta(w^a_{d_j}, w^a) \tag{2}$$

or

$$n(w^m|\nu, \mathbf{C} = c) = \sum_{j=1}^{\|\nu\|} \pi(w^a_{d_j}, \mathbf{C} = c)\delta(w^m_j, w^m) \tag{3}$$

where $\mathbf{C} = \{1, 2, .., C\}$ is the class label set. Eq. 2 and Eq. 3 indicate that the visual and motion information play predominant roles. $\pi(w^k_j, \mathbf{C} = c)$ is the attention information between feature w^k and class c. It will function as the weight of the other cue. For example, by Eq. 2, if motion is the predominant cue, we get an N-dimensional feature vector, where N is the number of visual words, and each element is a C-component of motion cue. Each component is the motion attention weight w.r.t class c. This attention based video representation indeed encodes both *what* and *how* aspects of an event. Each histogram is about the specific visual word which depicts *what* aspect, while the motion cue not only guides the impact of the visual word in capturing *how* aspect but also describes our prior knowledge about the categories we are looking for in the top-down manner. Similarly, by Eq. 2 the appearance information function as predominance cue is deployed to modulate the motion features. After concatenating the class-specific histogram, a video clip ν is eventually represented by a $N * C$-dimensional feature vector. In this paper, we propose two approaches to compute the attention information $\pi(w^k, \mathbf{C} = c)$: one is the probability of each word w^k w.r.t specific class c; the other is the mutual information between each word w^k and class c.

probabilistic vote We resort to the probability for every word w.r.t the specific class to characterize the impact of the local features on the video representation.

$$\pi(w^k, \mathbf{C} = c) = P(\mathbf{C} = c|w^k) \tag{4}$$

Given a visual or motion word, the probability of a class c $P(\mathbf{C} = c|w^k)$ can be estimated by using Bayes formula,

$$P(\mathbf{C} = c|w^k) = \frac{P(w^k|\mathbf{C} = c)P(\mathbf{C} = c)}{P(w^k)} \tag{5}$$

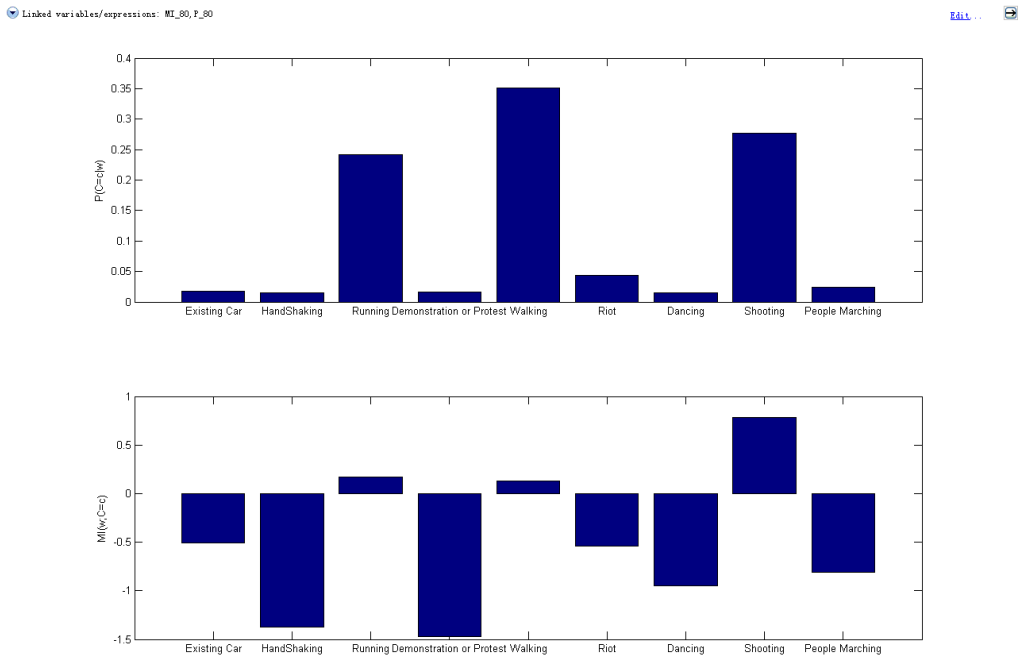

Fig. 2. Example class specific information given word w^k

where $P(w^k|\mathbf{C}=c)$ is the empirical distribution,

$$p(w^k|\mathbf{C}=c) = \frac{1}{\|\mathrm{T}^{c+}\|} \sum_{w^k_{d_j} \in \mathrm{T}^{c+}} \delta(w^k_{d_j}, w^k) \tag{6}$$

can be obtained by summing over the indexes to the positive training videos in class c. $P(w^k)$ is the probability of word w^k in all training videos.

$$p(w^k) = \frac{1}{\|\mathrm{T}\|} \sum_{w^k_{d_j} \in \mathrm{T}} \delta(w^k_{d_j}, w^k) \tag{7}$$

High probability w.r.t specific class c means more attention could be paid to the corresponding given word. However, if the probability is almost equal for every class, this word will be regarded as irrelevant for the recognition task. By Eq. 3, 2 and Eq. 4, Motion Probability based Appearance Histogram (MPAH) and Appearance Probability based Motion Histogram (APMH) can be obtained, in which motion and appearance cues are predominant respectively.

Mutual Information vote By Eq. 4 only the positive training information is take into consideration. In [17], better discriminative can be obtained by

incorporating the negative information. Therefore, we resort to the mutual information to measure the importance of the features. However, we evaluate the mutual information between a word and a specific class $c \in \mathbf{C}$ rather than the mutual information between a STIP and a specific class, since the latter needs the independence assumption.

$$\pi(w^k, \mathbf{C} = c) = MI(w^k; c) \tag{8}$$

where $MI(w^k, c)$ is the mutual information between word w^k and class c can be obtained by

$$\begin{aligned} MI(w^k; c) &= \log \frac{P(w^k|\mathbf{C} = c)}{P(w^k)} \\ &= \log \frac{P(w^k|\mathbf{C} = c)}{P(w^k|\mathbf{C} = c)P(\mathbf{C} = c) + P(w^k|\mathbf{C} \neq c)P(\mathbf{C} \neq c)} \\ &= \log \frac{1}{P(\mathbf{C} = c) + \frac{P(w^k|\mathbf{C} \neq c)}{P(w^k|\mathbf{C} = c)} P(\mathbf{C} \neq c)} \end{aligned} \tag{9}$$

From Eq. 9, we can see that the likelihood ratio test $\frac{P(w^k|\mathbf{C} \neq c)}{P(w^k|\mathbf{C} = c)}$ determines whether w^k votes positively or negatively for class c. If $\frac{P(w^k|\mathbf{C} \neq c)}{P(w^k|\mathbf{C} = c)} > 1$, then $MI(w^k, c) < 0$, which means this video word w^k votes a negative score for the class c. On the contrary, when $\frac{P(w^k|\mathbf{C} \neq c)}{P(w^k|\mathbf{C} = c)} < 1$, then $MI(w^k, c) > 0$,w^k votes a positive score for the class c. The likelihood $\frac{P(w^k|\mathbf{C} \neq c)}{P(w^k|\mathbf{C} = c)}$ can be obtained by

$$\frac{P(w^k|\mathbf{C} \neq c)}{P(w^k|\mathbf{C} = c)} = \frac{\frac{1}{\|\mathrm{T}^{c-}\|} \sum_{w^k_{d_j} \in \mathrm{T}^{c-}} \delta(w^k_{d_j}, w^k)}{\frac{1}{\|\mathrm{T}^{c+}\|} \sum_{w^k_{d_j} \in \mathrm{T}^{c+}} \delta(w^k_{d_j}, w^k)} \tag{10}$$

From the representation of $MI(w^k, c)$ we can observe that both positive and negative training information vote a score for the class c. Similarly, $MI(w^k, c)$ encodes how much information from word w^k in class c. High mutual information between word w^k and class label c means that the word feature w^k is highly relevant. Fig. 2 shows an example of the class specific information. For a motion word w^k extracted from TRECIVD 2005 video dataset, both the probability of each class and the mutual information w.r.t each class c is computed. Note that, generally, high probability corresponding to high information. In this instance, given a word, the probability of class "Walking"is the maximum, while the mutual information $MI(w^m, \mathbf{C} = 5)$ is positive thus means this word is relate to the video event with class label "Walking". However, the mutual information is not necessarily the highest. In contrast, the words with mutual information near zero are statistically independent from the class label, where the ones with negative mutual information vote a negative score for the corresponding label.

Some sample frames from the HOHA video dataset. From left to right: AnswerPhone, Kiss, SitUp, HandShake, SitDown, HugPerson, GetOutCar, StandUp

Some sample frames from the TRECVID 2005 video corpus. From left to right: Existing Car, Handshaking, Running, Demonstration or protest, Walking, Riot, Dancing, Shooting, People Marching.

Fig. 3. Example frames from HOHA and TRECVID dataset

4 Experiment

4.1 Data Sets

In order to demonstrate the performance of the proposed method, two different datasets :HOHA [5] and TRECVID 2005 [1], are used. Fig. 3 shows some sample pictures. HOHA contains 430 video clips, i.e., short sequences from 32 movies, of which 219 are used for training and 211 are used for testing. Each sample is annotated according to 8 classes: AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, StandUp. The ground truth of TRECVID 2005 dataset is based on LSCOM annotated events concepts [20]. After removing the events with a few positive samples, 9 of which are chosen as our evaluation set. Because the LSCOM annotation labels are deficient for our dynamic concepts, we re-annotated the event labels by watching all frames within the video shot. As a result, there are 1610 positive clips in total over the 9 events: Existing Car, Handshaking, Running, Demonstration Or Protest, Walking, Riot, Dancing, Shooting, People Marching, among which, half is used for classifier training and the remaining for testing. We evaluate the classification performance using the Average Precision (AP) measure, which is the standard evaluation metric employed in the TRECVID benchmark. Mean average precision (MAP) Average precision is proportional to the area under a recall-precision curve. To calculate AP for one concept, we first rank the test data according to the prediction of each sample. MAP is then calculated by averaging APs across all concepts.

4.2 Classifier

For classification, SVMs [21] are employed as "one-against-all"manner to estimate the likelihood of a given feature vector extracted from a video clip belonging to an event. In our experiments we use libSVM [21] with intersection kernel since

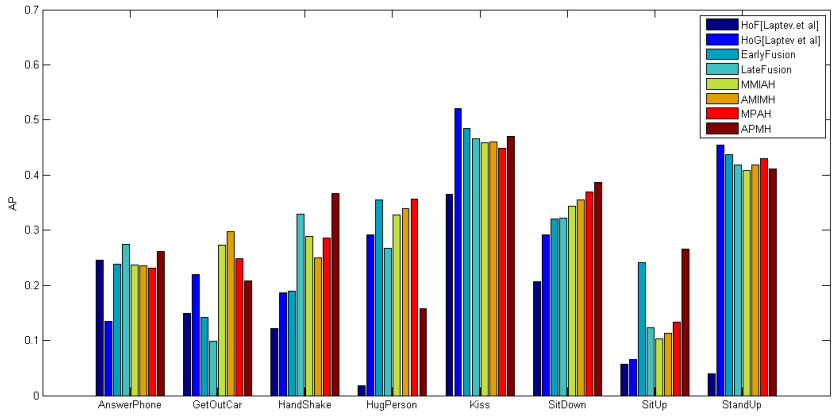

Fig. 4. Comparison of APs(%) between different features for recognizing action events in HOHA. HOG: Histogram of Oriented Gradient; HOF: Histogram of Flow; AMIMH: Appearance Mutual Information based Motion Histogram; MMIAH: Motion Mutual Information based Appearance Histogram; APMH: Appearance Probability based Motion Histogram; MPAH: Motion Probability based Appearance Histogram

it requires significantly less computational time while makes satisfactory results. For two BoW based histograms H_i and H_j extracted from video i and j, the intersection kernel is computed as:

$$K(H_i, H_j) = \frac{\sum_{n=1}^{N*C} \min(H_i(n), H_j(n))}{\min(\sum_{n=1}^{N*C} H_i(n), \sum_{n=1}^{n} H_j(n))} \tag{11}$$

4.3 HOHA

In this section, we present the results using our proposed method on HOHA dataset. We build a vocabulary of 600 visual words, 600 motions words and 600 combination words by clustering the HOG and HOF and HOG+HOF descriptors respectively. Fig. 4 shows the performances for eight actions on HOHA. Unlike most existing approaches which need object tracking, detection or ground substraction, our method is data driven and therefore does not need any preprocessing step. Moreover, there is no parameters needed to be determined in our method. From Fig. 4, we can observe that appearance (HOG) and motion (HOF) itself could not be a good guide to the performance of combined detector. Specially, Appearance Probability based Motion Histogram (APMH) achieves high mean AP (MAP) of 31.5%. This shows 17% improvement compared to HOG [5] (MAP=27%). Appearance Probability based Motion Histogram (APMH) outperforms EarlyFusion and LateFusion, this validates the proposed approach

Table 1. Comparison of Average Precision (%) using different features on TRECVID dataset

Event Name	HOG	HOF	HOG+HOF	LateFusion	MPAH	APMH	MMIAH	AMIMH
Existing-car	15.6	21.8	23.2	24.6	20.8	25.0	**30.3**	27.8
Handshaking	46.1	46.3	50.8	51.5	47.7	46.4	49.8	**54.2**
Running	76.0	76.7	78.3	76.3	76.6	77.6	76.4	**78.7**
Demonstration-Protest	26.4	25.0	16.3	25.2	26.2	26.8	**28.1**	27.5
Walking	72.0	71.2	72.3	**72.9**	72.1	72.0	70.9	72.2
Riot	28.9	27.2	28.0	28.2	30.1	28.8	30.6	**30.6**
Dancing	20.7	21.4	14.4	22.8	22.1	23.4	24.9	**26.3**
Shooting	60.1	65.6	65.3	62.2	61.8	65.1	**68.3**	65.7
People-Marching	21.9	20.6	22.6	21.0	21.1	21.9	**26.4**	24.1
Mean Average Precision	40.9	41.8	41.2	42.9	42.1	43.0	45.1	**45.2**

that the Appearance Probability based Motion Histogram does guide the action recognition. Specially, for some action events such as GetOutCar, HandShake, HugPerson, SitDown and SitUp the attention based combination feature methods perform best, and for the event "HandShake" mutual information vote perform best. It also can be seen that the improvement of mutual information based features in this dataset is limited. The reason is that mutual information is inclined to select rare words by Eq. refequ:MI2. On the other hand, the *what* aspect usually refers to a person in HOHA dataset, which means that the appearance features may not play a predominant role such that the appearance attention based motion histograms do not perform as good as motion attention based visual histograms.

4.4 TRECVID

We also quantitatively compare the event recognition accuracy by using the proposed algorithm with different features. Table. 1 presents the performances of nine events on TRECVID video corpus. As shown, we have a set of interesting observations:

1. Among these features, the best performance gain is obtained by Appearance Mutual Information based Attention (AMIMH) with the highest MAP of 45.2%, this shows 10.5% improvement compared with HOG (MAP=40.9%). The reason is that HOG only captures *what* aspect of an event, but ignores *how* aspect. Similarly, compared with HOF, an improvement of 8.13% is achieved in that HOF only captures *how* aspect, but ignores *what* ones. Both HoG+HoF and Late Fusion outperform HOG and HOF, which shows the necessity of the combination of these two cues. For the latter four features such as MPAH and MMIAH, appearance words capture *what* aspect, while their corresponding motion class specific histograms not only describe *how* aspect but also provide more relevant motion features for specific class.

2. In general, the attention based features outperform conventional multiple feature combination methods such as early fusion or late fusion strategies.

Unlike HOHA dataset, mutual information based attention approaches perform better than probability based attention in TRECVID dataset, which supports our argument in Section 3 that MI provides more discriminative features for the classification tasks by incorporating the negative votes of each word. The object are different from HOHA dataset, whereas former including person, car or other scenes.

3. A slight disappointing results of Motion Probability based Appearance Histogram compared with LateFusion method may be caused by the confusion motion probability of the given word. Such as for the event Existing car, the appearance of different cars are various, LateFusion has the property of "vocabulary compactness" [15], whereas the Motion Probability based Appearance Histogram lack it.

5 Conclusion

In this paper, we propose a novel approach by combining motion and visual features together for event recognition in Bag-of Words (BOW) framework. Given a visual or motion word, both the probability and mutual information of each class are used to guide the recognition in a top-down way. The results from TRECVID and HOHA dataset suggest that for most event categories attention based histograms not only capture two event aspects but also provide more discriminative features. Specially, no parameter needed to be determined within our approach.

References

1. http://www-nlpir.nist.gov/projects/trecvid.
2. Wang, F., Jiang, Y.G., Ngo, C.W.: Video event detection using motion relativity and visual relatedness. In: Proceeding of the 16th ACM international conference on Multimedia. (2008)
3. Xu, D., Chang, S.F.: Video event recognition using kernel methods with multilevel temporal alignment. IEEE Trans. Pattern Analysis and Machine Intelligence **30** (2008)
4. Zhou, X., Zhuang, X., Yan, S., Chang, S.F., Hasegawa-Johnson, M., Huang, T.S.: Sift-bag kernel for video event analysis. In: Proceeding of the 16th ACM international conference on Multimedia. (2008)
5. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)
6. Dollr, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking ans Surveillance. (2005) 65–72
7. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of ACM International Conference on Image and Video Retrieval. Volume 46. (2007)

8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. IEEE Conference on Computer Vision and Pattern Recognition **1** (2005) 886–893
10. Paul Scovanner, S.A., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceeding of the 15th ACM international conference on Multimedia. (2007)
11. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: International Conference on Computer Vision. (2005) 166 – 173
12. Ziming Zhang, Yiqun Hu, S.C., Chia, L.T.: Motion context: a new representation for human action recognition. In: Proceedings of the European Conference on Computer Vision. Volume 5305. (2008) 817–829
13. Rizwan Chaudhry, Avinash Ravichandran, G.H., Vidal, R. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)
14. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. (In: IEEE International Conference on Computer Vision) 726–733
15. Fahad Shabhaz Khan, Joost van de Weijer, M.V.: Top-down color attention for object recognition. In: International Conference on Computer Vision. (2009)
16. Liu, J., Shah, M.: Learning human actions via information maximization. In: IEEE Conference on Computer Vision and Pattern Recogntion. (2008) 1996–2003
17. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: IEEE Conference on Computer Vision and Pattern Recogntion. (2009) 2442–2449
18. Niebles, J.C., Wang, H., Fei-fei, L.: Unsupervised learning of human action categories using spatial temporal words. In: Proceedings of British Machine Vision Conference. (2006) 299–318
19. X.Chen, Zelinsky, G.J.: Real-world visual search is dominated by top-down guidance. Vision Research **46** (2006) 4118–4133
20. on Large Scale Concept Ontology for Multimedia, D.C.W.: Revison of lscom event/activity annotations (2006) Columbia University ADVENT Technical Report.
21. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.