Video Event Detection Using Motion Relativity and Feature Selection

Feng Wang, Zhanhu Sun, Yu-Gang Jiang, and Chong-Wah Ngo, *Member, IEEE*

***Abstract*—Event detection plays an essential role in video content analysis. In this paper, we present our approach based on motion relativity and feature selection for video event detection. First, we propose a new motion feature, namely Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW) to employ motion relativity for event detection. In ERMH-BoW, by representing *what* aspect of an event with Bag-of-Visual-Words (BoW), we construct relative motion histograms between different visual words to depict the objects' activities or *how* aspect of the event. ERMH-BoW thus integrates both *what* and *how* aspects for a complete event description. Meanwhile, we show that by employing motion relativity, ERMH-BoW is invariant to the varying camera movement and able to honestly describe the object activities in an event. Furthermore, compared with other motion features, ERMH-BoW encodes not only the motion of objects, but also the interactions between different objects/scenes. Second, to address the high-dimensionality problem of the ERMH-BoW feature, we further propose an approach based on information gain and informativeness weighting to select a cleaner and more discriminative set of features. Our experiments carried out on several challenging datasets provided by TRECVID for the MED (Multimedia Event Detection) task demonstrate that our proposed approach outperforms the state-of-the-art approaches for video event detection.**

***Index Terms*—Feature selection, motion relativity, video event detection.**

I. Introduction

With more and more multimedia data being captured to record the event occurrences in the real world and widely available from different sources such as the web, the management and retrieval of multimedia data has been actively researched in the past few decades, where multimedia content analysis serves as a fundamental and essential step. Content analysis of multimedia data, in nature, is event analysis, i.e. to detect and recognize events of user interest from different modalities such as video streams, audio and texts. A lot of efforts have been put to event-based video analysis including unusual event detection [3], [5], human action classification [12], [20], [26], [31], [27], [36], [40], and event recognition [14], [18], [23], [35], [48], [51].

(a) (b) (c)

Fig. 1. Difficulty in keyframe based event recognition. (a) *Airplane_Takeoff* or *Airplane_Landing*? (b) *Running*, *Dancing*, or *Walking*? (c) *Throwing* or *Catching*?

In the past decade, video semantic detection has attracted a lot of research attentions. In the video semantic indexing (SIN) task of annual TRECVID workshop [55], a benchmark of annotated video corpus is provided to researchers for detecting a set of predefined concepts. Besides the static concepts such as *Building* and *River*, some event-based concepts are also included, such as *Walking_Running* and *People-Marching*. Although great success has been achieved for video semantic detection, in the SIN task, researchers focus more on the detection of static concepts, while little attention has been paid to event detection.

Generally speaking, an event can be regarded as a semantic concept. However, in contrast to static concepts, event has its own nature, i.e. the dynamic nature. As a result, event detection is limited by the keyframe-based approaches that are widely used for the static concept detection. Without viewing the dynamic course of an event, human frequently encounter difficulties in event annotation. Fig. 1 shows the difficulty in keyframe-based event annotation and detection. For instance, in Fig. 1(a), by looking at the keyframe only, even for a human, it is difficult to judge whether the airplane is landing, taking off or just standing by in the lane. Event detection suffers from the incomplete representation of the keyframe for a dynamic event. Thus, in order to achieve better performance, it is necessary to employ the sequence information in event-based concept detection instead of the keyframe only.

Recently, more efforts have been paid to video event detection. Since 2010, TRECVID has provided a new task of multimedia event detection (MED) for researchers to evaluate their approaches [56]. Similar to video concept detection, MED is usually treated as a binary classification problem. Given a target event e, video clips are labelled as positive samples where e

Manuscript received June 20, 2013; revised November 09, 2013 and March 25, 2014; accepted March 25, 2014. Date of publication April 04, 2014; date of current version July 15, 2014. This work was supported in part by the grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 120213), the National Natural Science Foundation of China (No. 61103127 and 61375016), Shanghai Pujiang Program (No. 12PJ1402700), and the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shin'ichi Satoh.

F. Wang and Z. Sun are with the Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China (e-mail: fwang@cs.ecnu.edu.cn)

Y.-G. Jiang is with the School of Computer Science, Fudan University, Shanghai, China (e-mail: ygj@fudan.edu.cn).

C. W. Ngo is with the Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong (e-mail: cwngo@cs.cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2315780

Fig. 2. Motion relativity in event detection. Both two video clips contain the event *Walking*. Although camera movements are different during video capture, similar relative motion between *person* and *building* can be observed in both clips.

is present or negative ones where e is absent. SVM (Support Vector Machine) is widely adopted for the classification. Some attempts are made to adapt SVM to the sequential characteristics of event [48]. Different features that have proven effective in concept detection are also employed in event detection such as color, texture, audio, and local interest point (LIP) [22]. Compared with the keyframe based approach, the frame sequence is investigated during the feature extraction for event detection. Besides, volumetric and spatio-temporal features are also explored [17], [23], [47], [8], [46], [11], [38], [52].

In this paper, we focus on extracting effective features from video sequences for event detection. In a video clip, an event is usually described from two aspects: i) *what* are the objects/scenes that participate in the event, *e.g.* people, objects, buildings, etc.; ii) *how* the event evolves in temporal domain, i.e. the course of the event. The former consists of the static information and answers the questions like who, what, where, and when. These facets can basically be obtained from static images. The features to describe *what* aspect have been intensively studied, including global features (color moment, wavelet texture, edge histogram), local features (SIFT, ColorSIFT), and semantic features (concept score). The latter contains the dynamic information of the event and answers the question of *how*, *e.g.* the motion of the objects and the interactions among different objects/scenes. This information can only be captured by viewing the whole frame sequence. Motion is an important cue in describing the event evolution. Various motion features have been developed to capture the motion information in the sequence such as motion histogram [12] and motion vector map [18]. To completely describe an event, these two aspects should be closely integrated. For instance, Fig. 2 shows two video clips containing the event *Walking*. Intuitively, *"motion of person"* is important in describing *Walking*. *Person* and *Motion* are the two aspects of this event, which can be captured by the static features (e.g. color moment, SIFT) and the dynamic features (e.g. motion histogram) respectively. However, neither single one of them is enough to describe the event *Walking*.

In our preliminary work [44], we have pointed out another problem during the extraction of motion features in the video sequence, i.e. the observed motion in the video clip is distorted by the varying camera movement, and cannot depict the real object activities and interactions in an event. For instance, in the second clip of Fig. 2, the camera follows the person when he walks through the yard. No motion of the person can be detected by the traditional motion estimation. Therefore, the motion calculated in the frame sequence with reference to the moving cameras cannot honestly present the real activities of the person. However, in both two clips, we can see that the relative position between *Person* and the background scene (*Building*) is changing, and similar relative motion patterns can be consistently observed in different videos containing the same event. Thus, the relative motion is suitable to cope with camera movements for event description and detection.

Due to its ability to honestly describe the object activities in an event, in this paper, we employ motion relativity for event detection by proposing a new motion feature, namely Relative Motion Histogram of Bag-of-Visual-Words (RMH-BoW). Fig. 3 illustrates the procedure for our feature extraction. Considering that object segmentation and semantic annotation remains extremely difficult in unconstrained videos, we employ Bag-of-Visual-Words (BoW) with SIFT (Scale Invariant Feature Transform) which has been proven effective in concept detection to represent the presence of different objects/scenes, i.e. *what* aspect of an event. In BoW, a visual vocabulary is first constructed by grouping a set of local keypoint features using k-means. Then, a given video frame can be represented as a histogram of visual words by mapping its keypoints to the visual vocabulary.

With BoW capturing the objects in the videos, we then compute the motion of keypoints to estimate the motion of objects for event detection. This is based on the assumption that in different video samples, the objects of the same category should contain similar image patches/keypoints, and similar motion patterns of these keypoints could be observed if the same event is present. These patterns can thus be discovered and used to detect the event occurrences. In our approach, in order to eliminate the motion distortion caused by the camera movement, we employ the relative motion between visual words to capture the activities and interactions between different objects. For instance, as illustrated in Fig. 2, Q_{p1}, Q_{p2} and Q_{b1}, Q_{b2} are the keypoints lying on persons and buildings respectively. Although in both clips, the two persons walk in the similar way, Q_{p1} is moving while Q_{p2} remains still due to different camera movement. On the other hand, by investigating the relative motion between Q_{p1} and Q_{b1} in clip 1, and between Q_{p2} and Q_{b2} in clip 2, similar motion patterns can be observed to describe the motion relativity between *Person* and *Building* for detecting the

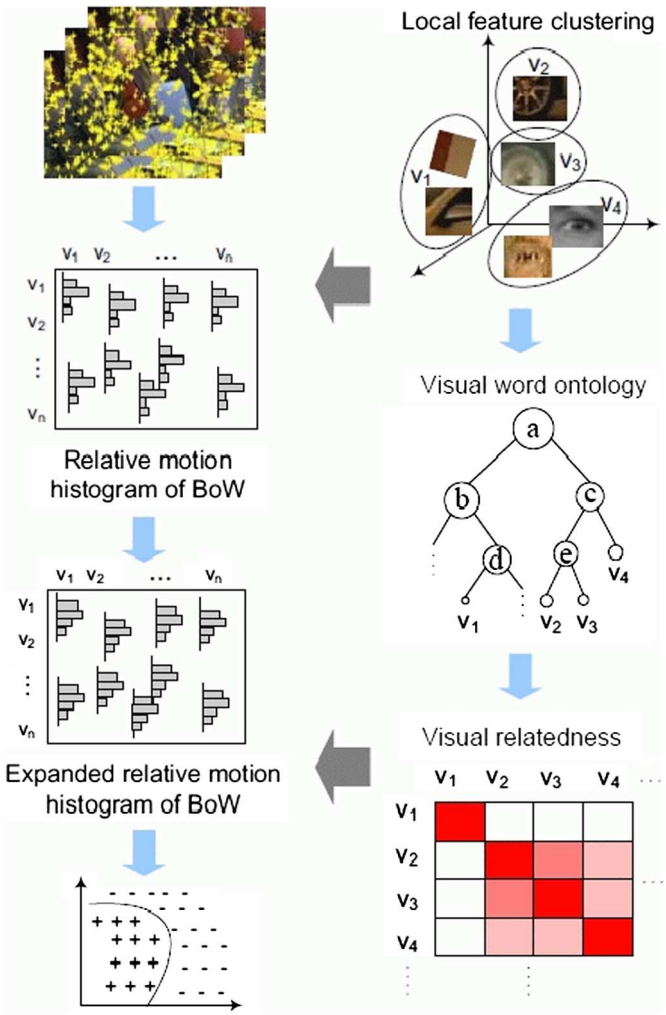


Fig. 3. Our proposed feature extraction framework for video event detection. With visual words capturing what are involved in an event, the local motion histogram of visual words describes both *what* and *how* aspects effectively for a complete representation of an event. Motion relativity and visual relatedness are employed to cope with the distortion by camera movement and the visual word correlation problem.

event *Walking*. As shown on the left of Fig. 3, given a video clip, the keypoints are tracked in neighboring frames and the relative motion is calculated between every two keypoints. Given two visual words, a relative motion histogram (RMH-BoW) is constructed by accumulating the motion vectors between every two keypoints mapped to the two words respectively. To alleviate the well-known visual word ambiguity problem [15] in BoW approach, we then employ visual relatedness between visual words [21] to expand the motion of a visual word to its nearest neighbors or correlated visual words to derive a new feature called Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW; detailed in Section III).

In summary, the effectiveness of ERMH-BoW in describing an event occurrence lies in three aspects: i) It closely integrates both the static (BoW with SIFT) and the dynamic information (motion) of an event in one feature; ii) It is invariant to the varying camera movement and thus able to discover the common motion patterns in the videos containing the same event; iii) It depicts not only the motion of the objects, but also the interactions between different objects/scenes which is important in describing event occurrences.

Since ERMH-BoW computes the relative motion between each pair of visual words, the dimensionality of the resulting feature is usually very large. This brings much difficulty to the feature storage and the classifier training. An event can usually be described by the interactions between few objects/scenes, and an object/scene can be captured by only a small set of keypoints or visual words. Thus, only very few elements in the ERMH-BoW feature are useful for detecting a target event. Based on this observation, in this paper, we extend our previous work in [44] by further proposing two approaches to select the informative features for ERMH-BoW in event detection so as to alleviate the curse of dimensionality. This is achieved by weighting the importance of the features in detecting different events and removing the less useful ones.

The remaining of this paper is organized as follows. Section II reviews some related works especially the existing features used in video event detection. Section III proposes and describes in detail the ERMH-BoW feature. In Section IV, we present the feature selection approach for reducing the dimensionality of the ERMH-BoW feature. Section V studies different strategies and settings in feature representation and event detection with the ERMH-BoW feature. Experimental results are presented in Section VI. Finally, Section VII concludes this paper.

II. Related Works

Many events can be represented as the object activities and interactions (such as *Walking* and *Airplane_Flying*), and show different motion patterns. Motion is thus an important cue in describing the course of an event. A great deal of efforts have been devoted to extracting effective features to capture the motion information in the video sequences. In some early works, motion features are used for video retrieval and classification. In [2], motion vectors extracted from MPEG compressed domains are used for video indexing. Segmentation and labeling are carried out based on motion vector clustering. Videos can then be indexed based on either global or segmentation features. In [30], a motion pattern descriptor namely motion texture is proposed for video retrieval and the classification of simple camera and object motion patterns. In [18], motion vectors are extracted from MPEG encoded videos and compressed to form a motion image. SVM is then used for event recognition. By experimenting on a small set of events, the feature is shown to be useful in recognizing events with different motion distribution patterns. In [13], spatio-temporal interactions between different objects are expressed by the predicate logic for video retrieval. This algorithm assumes the objects are correctly detected and located during video preprocessing. In [12], Motion History Image (MHI) is calculated over a frame sequence to describe the characteristic of the human motion. Recognition is achieved by statically matching MHIs. This approach is applied to well-segmented human figures for recognizing several predefined actions. In [23], an event is treated as a space-time volume in the video sequence. Volumetric features based on optical flow are extracted for event detection. This approach is used in videos

with single moving object (human) and action. In [39], a similarity measure is proposed to search for two different video segments with similar motion fields and behaviors. During the extraction of low-level visual features for event detection, motion has become the most important information to describe *how* an event evolves in the temporal dimension, e.g. the activities of objects and the interactions between different objects/scenes.

As discussed in Section I, neither single aspect (the static nor the dynamic information) can completely describe an event. Thus, it is important to integrate both aspects into one feature. Inspired by the success of local image features, in [25], Laptev and Lindeberg extend the notion of spatial interest points into the spatio-temporal domain and detect space-time interest point (STIP) as a compact representation of video data to reflect the interesting events. This is built on the idea of Harris and Forstner interest point operators and detect local structures in space-time domain where the image values have significant local variations in both space and time dimensions. Scale-invariant spatio-temporal descriptors are then computed to construct video representation in term of labeled space-time points and classify events. In [54], Everts *et al.* propose Color STIPs by further considering a number of chromatic representations derived from the opponent color space to improve the quality of intensity-based STIP detectors and descriptors. In [8], MoSIFT feature is proposed to capture the local information in both spatial and temporal domains. SIFT is employed for the spatial domain and the optical flow pyramid is used to compute the local motion. MoSIFT feature descriptors are constructed in the spirit of SIFT to be robust to small deformations through grid aggregation. In [46], inspired by dense sampling in image classification, dense trajectories are extracted by first sampling dense points from each frame and then tracking them based on displacement information from a dense optical flow field. Descriptors are computed based on motion boundary histograms which is robust to camera motion. In [11], the orientated histograms of differential optical flow are used for motion coding and combined with the histogram of oriented gradient appearance descriptors to detect standing and moving people in videos with moving cameras and backgrounds. These features encode both the appearance information in spatial domain and the motion information in temporal dimension. According to the evaluation reported in [41], they have demonstrated to be robust for event detection in open video sources and achieved encouraging results.

For feature representation, Bag-of-Words (BoW) approach has demonstrated to be surprisingly effective for most existing features. In BoW approach, a feature vocabulary is first constructed by quantizing the descriptor space with clustering. A BoW histogram is then computed by counting the presence of different words in the video volume to represent the video content. In [22], BoW approach is used to encode different features extracted from both static frames and video sequences such as SIFT, spatio-temporal interest points and MFCC. In BoW approach, event occurrences are detected based on the presence of some specific static or motion features. Besides BoW approach, in some other works [14], [48], an event is viewed as a temporal process over the feature spaces. Features are first extracted from each keyframe (or sub-clip) of a given video, and the video is then represented as a sequence of feature vectors to encode the static/motion information at each moment along the temporal dimension. In [14], visual events are viewed as stochastic temporal processes in the semantic space. The dynamic pattern of an event is modeled through the collective evolution patterns of the individual semantic concepts in the course of the visual event. HMM (Hidden Markov Model) is employed for event modeling and recognition. In [53], a long-duration complex activity is decomposed into a sequence of simple action units. Probabilistic suffix tree is proposed to represent the Markov dependencies between the action units. In [48], a video clip is represented as a bag of descriptors from all of the constituent frames. EMD (Earth Mover's Distance) is applied to integrate similarities among frames from two clips, and TAPM (Temporally Aligned Pyramid Matching) is used for measuring the video similarity. EMD distance is then incorporated into the kernel function of SVM framework for event detection. This kind of approach aims at discovering patterns of event evolution along the temporal dimension.

III. Expanded Relative Motion Histogram of BoW (ERMH-BoW)

In this section, we propose to employ motion relativity for developing an effective feature, namely Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW) for video event detection. In our approach, Bag-of-Visual-Words (BoW) features are employed to capture *what* aspect of an event. For the construction of the visual vocabulary, keypoints are detected on a set of training images using Difference of Gaussian (DoG) [28] and Hessian-Laplacian detectors [33], and described with SIFT [28]. K-means algorithm is then employed to cluster all keypoints and construct a visual vocabulary, where each cluster is treated as a visual word. Given a keyframe extracted in the video, by mapping the detected keypoints to the visual vocabulary [19], we can represent the content of the keyframe as a vector of visual words with its weights indicating the presence or absence of the visual words.

With BoW capturing *what* is present in a given video, we then extract the motion information to capture *how* an event evolves along the video sequence. First, we construct a motion histogram for each visual word (MH-BoW) by capturing the local motion information of the keypoints. Second, to employ motion relativity, we modify MH-BoW by replacing the motion vectors with the relative motion between different visual words (RMH-BoW). Finally, to alleviate the mismatch problem [15], [21] existing in the BoW approach, we expand the motion histogram by considering the correlations between different visual words (ERMH-BoW).

A. Motion Histogram of Visual Words (MH-BoW)

In this section, we construct a local motion histogram for each visual word in BoW. To be efficient, 5 keyframes are evenly sampled every second in the video. Our motion features are extracted between every two neighboring keyframes. Given a keyframe, keypoints are first detected by DoG [28] and Hessian-Laplacian detectors [33]. We then employ the algorithm in [29] to track the keypoints in the next keyframe. For each keypoint r that can be successfully tracked, we calculate its motion vector m_r between these two frames. Different from

other motion histograms that are the sums of motion vectors over spatial regions, our motion histogram of BoW (MH-BoW) is constructed by summing up motion vectors of all keypoints mapped to the same visual word. For each visual word, we construct a 4-directional histogram. For this purpose, the motion vector m_r is decomposed into four components $D_i(m_r)$, where $i = 1, 2, 3, 4$ are corresponding to the four directions: left, right, up and down, and $D_i(.)$ projects m_r to the $i-th$ direction. For a visual word v, the motion histogram is calculated as

$$H_i(v) = \sum_{p \in N_v} D_i(m_r), \quad i = 1, 2, 3, 4 \tag{1}$$

where N_v is the set of tracked keypoints that are mapped to the visual word v.

By Equation (1), we get an S-dimension feature vector called Motion Histogram of BoW (MH-BoW), where S is the size of the visual vocabulary, and each element is a 4-directional motion histogram for the corresponding visual word. MH-BoW indeed encodes both *what* and *how* aspects of an event in a single feature. Each histogram is corresponding to a specific visual word which describes *what* aspect, while the motion histogram depicts the motion pattern and intensity of the visual word to capture *how* aspect. Since the local motion of the visual words is employed in MH-BoW, different events can be represented as certain motion patterns of specific visual words depicting different objects.

B. Relative Motion Histogram between Visual Words (RMH-BoW)

In MH-BoW, the motion is calculated as the movement of keypoints with reference to the camera which can be easily distorted by the camera movement as illustrated in Fig. 2. Furthermore, it just captures the motion of each isolated object and ignores the interaction between different objects/scenes which is important in describing an event occurrence. To address these problems, we propose to employ motion relativity between different objects and scenes for event detection. As discussed in Section I and observed in Fig. 2, motion relativity remains consistent for different clips containing the same event regardless of varying camera movement. In other words, it is able to honestly describe the real object activities and interactions in an event. Based on this observation, we modify MH-BoW in Section III-A with the relative motion histograms between visual words.

Given two visual words a and b, the relative motion histogram between them is calculated as

$$R_i(a, b) = \sum_{r \in N_a, t \in N_b} D_i(m_r - m_t) \tag{2}$$

where r and t are keypoints mapped to visual words a and b respectively, $m_r - m_t$ is the relative motion of r with reference to t, and $D_i(.), i = 1, 2, 3, 4$ decomposes the relative motion vector to the four directions as in Section III-A to generate a 4-directional histogram between visual words a and b. By Equation (2), the motion information in a video clip is represented as an $S \times S$ matrix R, where each element $R(a, b)$ is a relative motion histogram between the two visual words a and b. We call this feature matrix Relative Motion Histogram of BoW (RMH-BoW).

As seen in the derivation process, RMH-BoW depicts the intensities and patterns of relative motion between different visual words. Since the visual words in BoW capture *what* aspect of an event, RMH-BoW can be used to describe the activities and interactions between different objects and scenes in an event. Intuitively, different events are presented as different object motion patterns and intensities, while video clips containing the same event show similar motion patterns and intensities between specific objects or scenes. RMH-BoW can thus be used in supervised learning to discover these common patterns in different clips containing the same event for effective detection.

C. Expanding RMH-BoW with Visual Word Relatedness (ERMH-BoW)

When BoW is used to represent *what* aspect of an event, some visual words may be correlated (i.e. depicting the same object category), but are treated as isolated to each other [15], [21]. This will cause feature mismatch problem between events containing the same object. In this section, to address the visual word correlation problem in RMH-BoW, we expand the relative motion histogram based on visual relatedness. The expansion is conducted by diffusing the motion histograms across correlated visual words.

C.1 Visual Relatedness: The visual relatedness is a measurement of visual word similarity. In this paper, we employ the approach in our previous work [21] to estimate the relatedness between different visual words in a similar way as we estimate the semantic relatedness of textual words using general ontology such as WordNet. Based on the visual vocabulary, a visual ontology is further generated by adopting agglomerative clustering to hierarchically group two nearest visual words at a time in the bottom-up manner. Consequently, the visual words in the vocabulary are represented in a hierarchical tree, namely visual ontology, where the leaves are the visual words and the internal nodes are ancestors modeling the *is-a* relationship of visual words. An example of the visual ontology is shown on the right of Fig. 3. In the visual ontology, each node is a hyperball in the keypoint feature space. The size (number of keypoints) of the hyperballs increases when traversing the tree upward.

Similar to the semantic relatedness measurements of text words, the visual relatedness can also be estimated by considering several popular ontological factors based on the visual ontology. We directly apply a text linguistic measurement, JCN, to estimate the visual relatedness. Denote a and b as two visual words, JCN considers the ICs (Information Content) of their common ancestor and the two compared words, defined as:

$$\mathrm{JCN}(a, b) = \frac{1}{\mathrm{IC}(a) + \mathrm{IC}(b) - 2 \cdot \mathrm{IC}(\mathrm{LCA}(a, b))} \tag{3}$$

where LCA is the lowest common ancestor of visual words a and b in the visual ontology. IC is quantified as the negative log likelihood of word/node probability:

$$IC(a) = -\log p(a) \tag{4}$$

where the probability $p(a)$ is estimated by the percentage of keypoints in the visual hyperball a.

C.2 Expanding RMH-BoW: Based on the visual relatedness calculated by JCN, we expand RMH-BoW by diffusing the relative motion histograms between two visual words to their correlated visual words. The Expanded Relative Motion Histogram of BoW (ERMH-BoW) is calculated as

$$E(a,b) = R(a,b) + \sum_{s_a, s_b} \mathrm{JCN}(s_a, a) \times R(s_a, s_b) \times \mathrm{JCN}(s_b, b) \tag{5}$$

where $\{s_a\}$ and $\{s_b\}$ are the sets of visual words that are correlated to the words a and b respectively. The aim of RMH-BoW expansion is to alleviate the problem of visual word correlation. More specifically, the relative motion between two words are diffused by the influence of other words that are ontologically related to them. The diffusion inherently results in the expansion of RMH-BoW to facilitate the utilization of word-to-word correlation for video clip comparison. For instance, in Fig. 2, if the two keypoints Q_{p1} and Q_{p2} lying on *Person* are assigned to different visual words, say v_1 and v_2 respectively, this will cause mismatch in RMH-BoW. With ERMH-BoW, given that v_1 and v_2 are highly correlated, their corresponding motion histograms will be diffused to each other, and thus can be matched with higher similarity as expected. In our experiments, for each visual word, we empirically choose the five most similar words for diffusion in Equation (5). On one hand, this guarantees the efficiency of the RMH-BoW expansion process; on the other hand, diffusing with more visual words does not promise better performance.

Similar to BoW feature representations, we call each element in the ERMH-BoW feature a *motion word* which encodes the relative motion pattern between two visual words. The resulting ERMH-BoW feature counts the presence of different motion words in the given video clip. In the following sections, for the ease of presentation, we denote the ERMH-BoW feature matrix with a feature vector by concatenating all rows in the matrix together. Given a video p, the resulting feature vector is represented as $x_p = (x_{p1}, x_{p2}, \cdots, x_{pn})$.

IV. Feature Selection for ERMH-BoW

In RMH-BoW, to employ the motion relativity for event description, the relative motion between each pair of visual words is computed. This results in a sparse matrix with a high dimensionality. By word expansion, the problem becomes even worse since the sparsity of the feature is reduced. Eventually, it is space-consuming to store the ERMH-BoW features and time-consuming to train and test classifiers due to the curse of dimensionality. In this section, we reduce the high dimensionality of ERMH-BoW by removing the less useful motion words.

Feature selection has been intensively studied in text categorization. In video semantic indexing, similar attempts have been made to capture the concept-specific visual information by selecting the informative visual words and removing the useless ones [24], [43]. The ERMH-BoW feature actually captures all possible motion patterns between different visual words. However, an event can usually be presented as certain activities or interactions between few objects/scenes, which can be captured by a small set of visual and motion words. Thus, many words in ERMH-BoW feature are useless and even noisy in detecting given events. Based on this observation, we reduce the dimensionality of ERMH-BoW feature by measuring the importance of motion words for event detection and removing the less useful ones. In our approach, a two-step approach is employed. First, we filter those words that are useless in discriminating different event classes and generating a universal set of words for all events. This is achieved by employing the information gain approach used in text categorization to measure the discriminative ability of each word. Second, in the remaining words, we then measure the informativeness of each word for detecting a given event. This results in a specific set of words for each event.

A. Information Gain

Information gain (IG) is frequently employed to measure the goodness of features in machine learning, and has shown to be one of the best approaches for feature selection in text categorization [49]. In our approach, we employ IG to select motion words that are important in discriminating different events. Let $\{e_i\}$ denote the set of categories in event detection. The information gain of a word v is defined to be

$$\begin{aligned} G(v) = &- \sum_i P(e_i) \log P(e_i) \\ &+ P(v) \sum_i P(e_i|v) \log P(e_i|v) \\ &+ P(v) \sum_i P(e_i|\bar{v}) \log P(e_i|\bar{v}) \end{aligned} \tag{6}$$

In Equation (6), the IG value of v measures the information obtained for categorizing different events by knowing the presence or absence of v in a video. In the training video corpus, we compute the information gain of each word in ERMH-BoW and remove those with the IG values lower than a threshold. The determination of the threshold is discussed in our experiments (Section VI-A).

B. Event-Specific Feature Selection

With information gain, we mainly remove the motion words that are useless or even noisy in categorizing different event classes. The computation of information gain is relatively fast. However, feature selection is disconnected from the classifier learning process. Furthermore, IG approach generates a universal set of words for the detection of all events. Actually, different motion words are not equivalently important in detecting a specific event. A word which is important for detecting one event may not be useful in the detection of another event. The similar problem has been discussed and addressed in video concept detection by constructing concept-specific visual vocabulary. In [43], we propose to weight the informativeness of visual words for detecting specific concepts by SVM kernel optimization. In this paper, we revise the approach in [43] for feature selection of ERMH-BoW.

B.1 Problem Formulation: Event detection is usually treated as a binary classification problem, where SVM is widely adopted. The performance of SVM is largely dependent on its

kernel. Here we only consider the detection of a specific event e. As discussed above, different motion words are not equally important for detecting event e. To measure the importance of motion words, we assign different weights to them. Here we take $\chi^2 - RBF$ kernel as an example for discussion. Similar approach can be easily applied to other kernels. Originally $\chi^2 - RBF$ kernel is defined as

$$K_{pq} = exp(-\sigma \cdot d(x_p, x_q)) \tag{7}$$

$$d(x_p, x_q) = \sum_{i=1}^{n} \frac{(x_{pi} - x_{qi})^2}{x_{pi} + x_{qi}} \tag{8}$$

where x_p, x_q are the feature vectors for two video samples p and q. As can be seen in Equations (7) and (8), all motion words are treated equally for detecting different events. By assigning different weights to motion words, Equation (8) is rewritten as

$$d(x_p, x_q) = \sum_{i=1}^{n} w_i \cdot \frac{(x_{pi} - x_{qi})^2}{x_{pi} + x_{qi}} \tag{9}$$

where w_i measures the importance of the i-th word for detecting event e. An optimal weight vector $w = (w_1, w_2, \cdots, w_n)$ can be estimated by maximizing the discriminative ability of the SVM kernels.

In this paper, we employ the Kernel Alignment Score (KAS) [10] to measure the discriminative ability of SVM kernel, which is defined as

$$\bar{T} = \frac{\sum_{p,q} K_{pq} \cdot l_p \cdot l_q}{N \cdot \sqrt{\sum_{p,q} K_{pq}^2}} \tag{10}$$

where l_p is the label of p, $l_p = +1$ (or -1) if p is a positive (or negative) sample, and N is the total number of samples. The KAS score computed by Equation (10) measures how well an actual kernel is aligned with an optimal kernel [10] in which the distance between samples of different classes should be maximized and the distance between samples of the same class should be minimized. Generally, a kernel with higher KAS score is better at discriminating samples of different classes, and can potentially achieve better performance for classification.

Equation (10) assumes the two classes are balanced. However, this is not the case for most current datasets in video event detection, where there are usually many more negative examples than positive ones. This may bias the resulting KAS towards the negative class. To deal with this imbalance problem of the datasets, in [43], we modify Equation (10) by assigning different weights to the positive and negative examples as follows

$$\alpha_p = \begin{cases} 1 & \text{if } l_p = -1 \\ \frac{N^-}{N^+} & \text{otherwise} \end{cases} \tag{11}$$

where N^- and N^+ are the numbers of negative and positive examples in the training dataset respectively. Equation (10) is then modified as

$$T = \frac{\sum_{p<q} K_{pq} \cdot l_p \cdot l_q \cdot \alpha_p \cdot \alpha_q}{N' \cdot \sqrt{\sum_{p<q} \alpha_p \cdot \alpha_q \cdot K_{pq}^2}} \tag{12}$$

where $N' = \sum_{p<q} \alpha_p \cdot \alpha_q$. Eventually the problem of informativeness weighting for motion words is formulated as searching for an optimal weight vector w^{opt} such that the KAS score T defined by Equation (12) is maximized.

B.2 Gradient-based Weight Optimization: In our approach, we weight the importance of motion words by adopting a gradient-descent algorithm to maximize the KAS score in Equation (12). We calculate the partial derivative of T to the weight w_i as

$$\frac{\partial T}{\partial w_i} = \sum_{p<q} \frac{\partial T}{\partial K_{pq}} \cdot \frac{\partial K_{pq}}{\partial w_i} \tag{13}$$

$$\frac{\partial K_{pq}}{\partial w_i} = K_{pq} \cdot \left(-\sigma \cdot \frac{\partial d(x_p, x_q)}{\partial w_i} \right) \tag{14}$$

Based on Equations (13) and (14), we iteratively update the weight vector w of motion words so as to maximize the kernel alignment score defined by Equation (12). Below is the algorithm for optimization:

1) *Initialize $w_i = 1$ for $i = 1, 2, \cdots, n$. Calculate the initial KAS score T by Equation (12).*
2) *Update weights $w_i' = w_i + sign\left(\frac{\partial T}{\partial w_i}\right) \cdot \delta_w$, where* $sign(t) = \begin{cases} +1 & if \quad t > 0 \\ 0 & if \quad t = 0 \\ -1 & if \quad t < 0 \end{cases}$.
3) *Calculate the new kernel alignment score T' using the updated weights. If $\frac{T'-T}{T} < thres$, stop; otherwise, $T = T'$ and go to step 2.*

In step 2, the weight vector is updated by a value δ_w for each iteration. In our implementation, to be efficient, δ_w is set to be $\frac{1}{8}$, i.e. stepwise weights are used for different motion words. After the optimization process, we remove those motion words with weights less than a threshold which is empirically determined in Section VI-A. Finally, an event-specific set of motion words are selected. By feature selection, the dimensionality of ERMH-BoW feature is reduced, while the discriminative ability of SVM is improved.

V. Video Event Detection with ERMH-BoW

In this section, we employ the proposed feature ERMH-BoW for video event detection. In most existing systems, two different approaches are employed for feature representation. The first one is Bag-of-Word approach, where a feature vocabulary is constructed and a single feature vector is then extracted as the histogram on the vocabulary to present the content of the whole video. For the second approach, sequential information is employed. A feature vector is extracted from each keyframe (or sub-clip) and a sequence of vectors are used to represent the evolution of the event over the feature space along the timeline. In this paper, we compare these two different feature representations for event detection with ERMH-BoW. Given an event, an SVM is trained to classify positive and negative video samples. For two different feature representations, $\chi^2 - RBF$ kernel and EMD (Earth Mover's Distance) kernel proposed in [48] are employed respectively.

A. RBF Kernel

In this approach, the whole video is treated as a volumetric object for feature extraction. Five keyframes are evenly sampled every second. ERMH-BoW features are extracted between neighboring keyframes and then accumulated over the whole video clip. The resulting feature vector mainly encodes the motion patterns and intensities between different visual words in the video. Finally, $\chi^2 - RBF$ kernel in Equation(7) is used for SVM classification. The advantage of this approach is its efficiency since only one feature vector is extracted over the whole clip. However, it just captures limited evolution information of an event, i.e. the motion between neighboring keyframes.

B. EMD Kernel

To capture more sequence information in the video, we first represent a given video as a sequential object by evenly segmenting it into fixed-length sub-clips. An ERMH-BoW feature vector is then extracted from each sub-clip as described in Section V-A. This eventually results in a sequence of vectors to capture how an event evolves along the timeline in the video. For event detection, we employ the approach proposed in [48] to measure the similarity between different videos with EMD (Earth Mover's Distance) and then incorporate it into SVM kernel for classification.

For video clip similarity measure, EMD has proven to be effective in video clip alignment and matching. To employ EMD for video clip similarity measure, the ground distance between a pair of sub-clips from two videos is defined as the Euclidean distance of the ERMH-BoW features corresponding to the two sub-clips:

$$d(x_a, x_b) = \sqrt{\frac{1}{n} \sum_{1 \le i \le n} (x_{ai} - x_{bi})^2} \tag{15}$$

Given two videos $p = \{(x_p^{(1)}, w_p^{(1)}), (x_p^{(2)}, w_p^{(2)}), \cdots, (x_p^{(k)}, w_p^{(k)})\}$ and $q = \{(x_q^{(1)}, w_q^{(1)}), (x_q^{(2)}, w_q^{(2)}), \cdots, (x_q^{(l)}, w_q^{(l)})\}$ with k and l ERMH-BoW vectors as signatures respectively, and $w_p^{(i)} = 1/k$, $w_q^{(j)} = 1/l$, $1 \le i \le k, 1 \le j \le l$ as the weights for each ERMH-BoW vector. The EMD distance between p and q is computed by

$$D(p, q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d(x_p^{(i)}, x_q^{(j)})}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{16}$$

where f_{ij} is the optimal match among two sequences of ERMH-BoW vectors of p and q. The details of how to determine f_{ij} can be found in [48].

With the EMD distance between video samples computed in Equation (16) by employing ERMH-BoW features, we adopt the algorithm in [48] to train SVMs for event detection. The EMD distance between videos is incorporated into the kernel function of the SVM framework by using Gaussian function:

$$\hat{K}_{pq} = exp(-\frac{1}{\kappa M} D(p, q)) \tag{17}$$

where the normalization factor M is the mean of the EMD distances between all training videos, and κ is a scaling factor empirically decided by cross-validation.

In [48], the positive definiteness of the EMD kernel has been verified by experiments. Pyramid matching with different levels is fused to achieve better results. In this work, since we focus on feature extraction for event description, we aim at validating the effectiveness of the proposed feature ERMH-BoW. To be efficient, we just adopt a single-level EMD matching algorithm for distance measure.

VI. Experiments

In this section, we conduct experiments to compare the proposed ERMH-BoW feature with existing features to validate the effectiveness of motion relativity and feature selection for video event detection. We use the data resources provided by NIST in the yearly MED (Multimedia Event Detection) task [56] for experiment. The following datasets are included.

- *MED10 data*: This dataset consists of 3468 video clips with a total duration of about 114 hours. The data was collected by the Linguistic Data Consortium, and consists of publicly available, user-generated content posted to various Internet video hosting sites.
- *MED11 Transparent Development (DEVT) collection*: This collection includes 10403 video clips with a total duration of about 324 hours and was used for system development in MED11.
- *MED11 Opaque Development (DEVO) collection*: The DEVO corpus contains 32061 video clips with a total duration of about 991 hours and was used for system evaluation in MED11.
- *Progress Test collections (PROGTEST)*: This dataset is constructed by the Linguistic Data Consortium and NIST, and contains 98117 video clips with a total duration of about 3722 hours as the test set in MED12 task. The groundtruth of this dataset is currently not available.
- *Event kits*: Twenty events which are evaluated in MED12 are used in our experiment. The detailed definitions and information of the events can be found at [56]. In total the event kits contain 3736 example videos for the events. The number of positive examples for each event ranges from 119 to 221 as shown in Table I.

All video clips in the above datasets are provided in MPEG-4 formatted files with the video being encoded to the H.264 standard and the audio being encoded using MPEG-4's Advanced Audio Coding (ACC) standard [56]. We present our experimental results on MED12 development set and test set respectively. For the former, we take positive examples from the event kits provided by MED 12 (see Table I). Half of them are used for training and another half for testing. The negative examples are taken from MED10, DEVT and DEVO collections by removing those positive clips containing any one of the twenty events according to the groundtruth annotations from MED12. We use 30% of negative examples for training and the other 70% for testing. For the experiments on the test set, since the groundtruth of PROGTEST collection is strictly protected and blind to researchers until 2015, we only present and discuss the official evaluation results in our participation in MED 12 where all positive examples in event kits are used for training and the PROGTEST collection for testing.

TABLE I
EVENT KITS USED IN MED12

Event ID	Event Name	# positive
E006	Birthday party	221
E007	Changing a vehicle tire	119
E008	Flash mob gathering	191
E009	Getting a vehicle unstuck	151
E010	Grooming an animal	143
E011	Making a sandwich	186
E012	Parade	171
E013	Parkour	134
E014	Repairing an appliance	137
E015	Working on a sewing project	124
E021	Attempting a bike trick	200
E022	Cleaning an appliance	200
E023	Dog show	200
E024	Giving directions to a location	200
E025	Marriage proposal	200
E026	Renovating a home	200
E027	Rock climbing	200
E028	Town hall meeting	200
E029	Winning a race without a vehicle	200
E030	Working on a metal crafts project	200

In our implementation, the detection of a given event e is treated as a *one vs. all* binary classification problem and the system outputs a confidence score for each video. We employ the DET Curve [32], [57] to evaluate the performance of event detection. The DET Curves involve a tradeoff of two error types: Missed Detection (MD) and False Alarm (FA) errors, which are defined as

$$P_{MD}(e, thres) = \frac{\#MD(e, thres)}{\#Targets(e)} \tag{18}$$

$$P_{FA}(e, thres) = \frac{\#FA(e, thres)}{\#TotalClips - \#Targets(e)} \tag{19}$$

where *thres* is a threshold to determine whether a video contains event e, $\#MD(e, thres)$ is the number of missed detection (positive clips with confidence scores lower than *thres*), $\#FA(e, thres)$ is the number of false alarms (negative clips with confidence scores higher than *thres*), $\#Targets(e)$ is the number of positive clips in the groundtruth, and $\#TotalClips$ is the total number of clips in the test set.

Fig. 4 shows an example of DET Curve with different features for detecting event *Parkour* on the development set. Based on DET Curve, to numerically compare the performances of different approaches, we present P_{MD} values when $P_{FA} = 5\%$. This threshold setting was used in [41] for evaluating the performances of different features. Although this may cause many arguments on the selection of the P_{FA} threshold, it can demonstrate the performances of different approaches by looking at the detection accuracy at a fixed point considering that most people focus more on a relatively small set of returned results with a higher precision. In the following, we present the evaluation results on the development set and the MED12 Progress Test Collections respectively.

A. On Development Set

A.1 Performance of ERMH-BoW Feature: Table II compares the performances of different approaches for the detection of 20 events on the development set. For the extraction of MH-BoW, RMH-BoW and ERMH-BoW features, each video is equally segmented into a number of 5-second sub-clips, and a motion histogram is computed over each sub-clip. This results in a sequence of histograms for each video and EMD kernel is used for SVM training and testing. To compare the performances of two feature representations described in Section V, in another approach (ERMH-RBF), we also extract ERMH-BoW features over the whole video sequence to compose a single vector and employ $\chi^2 - RBF$ kernel for SVM classification. Furthermore, we compare our approaches with three motion features: MoSIFT [8], DTF-HoG [46], and DTF-MBH [11]. According to the evaluation results reported by [41], MoSIFT and DTF-HoG perform the best among different low-level features for event detection. For the extraction of MoSIFT, DTF-HoG, and DTF-MBH features, we follow the approaches presented in [41].

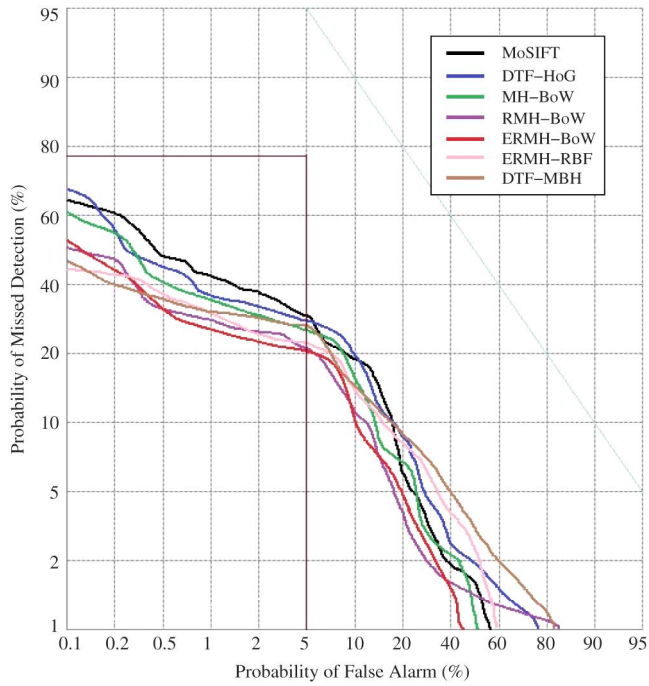


Fig. 4. DET Curve of different features for event E013 (Parkour).

As can be seen in Table II, on this dataset, MoSIFT performs better than DTF-HoG and DTF-MBH. The performances of these three features are basically consistent with the evaluation results in [41]. On average, MH-BoW achieves similar performance with MoSIFT. Actually they capture similar information in videos, i.e. the local motion information of keypoints. This shows that a simple representation of the motion information can already achieve competitive results for the detection of many events such as *Flash_mob_gathering*, *Repairing_an_appliance* and *Winning_a_race_without_a_vehicle*. Meanwhile, the motion features do not work well enough for some events such as *Grooming_an_animal* and *Making_a_sandwich*. The possible reason mainly lies in two aspects: i) Some events are not motion intensive and do not show strong and consistent motion patterns; ii) The representation of the motion information could be further improved to derive more effective features.

TABLE II
COMPARISON OF DIFFERENT APPROACHES ON THE DEVELOPMENT SET. THE PERFORMANCES ARE EVALUATED BY P_{MD} VALUES WHEN $P_{FA} = 5\%$

Events	MH-BoW	RMH-BoW	ERMH-BoW	ERMH-RBF	MoSIFT	DTF-HoG	DTF-MBH
Birthday_party	**31.56**	33.88	34.34	33.05	35.62	40.17	50.31
Changing_a_vehicle_tire	37.43	34.15	**32.03**	35.64	38.39	42.28	41.09
Flash_mob_gathering	17.71	**13.68**	13.94	16.23	14.74	21.36	25.44
Getting_a_vehicle_unstuck	37.62	30.74	**27.65**	29.88	34.75	30.43	35.62
Grooming_an_animal	44.14	45.4	**43.17**	45.62	48.81	46.53	49.55
Making_a_sandwich	47.35	49.22	44.73	**43.37**	43.5	48.94	52.16
Parade	22.08	17.93	**17.21**	19.86	24.11	30.65	34.79
Parkour	27.95	21.36	**20.16**	23.79	31.25	30.86	28.24
Repairing_an_appliance	18.42	15.56	15.98	17.15	**14.83**	27.47	31.88
Working_on_a_sewing_project	34.27	30.69	31.1	32.11	**29.47**	35.53	37.9
Attempting_a_bike_trick	25.88	19.72	**18.84**	19.91	22.47	28.58	29.28
Cleaning_an_appliance	23.61	**20.39**	20.45	22.14	20.52	25.95	26.74
Dog_show	20.45	16.28	**15.46**	15.98	17.66	27.74	30.18
Giving_directions_to_a_location	47.21	49.53	44.61	43.51	46.79	**42.15**	52.37
Marriage_proposal	46.75	**41.93**	42.61	46.06	44.83	49.24	48.26
Renovating_a_home	20.1	17.45	**16.09**	17.63	19.4	23.29	21.85
Rock_climbing	24.36	**19.67**	20.43	22.26	28.39	24.37	26.66
Town_hall_meeting	52.18	50.21	47.39	48.41	52.15	**46.95**	57.59
Winning_a_race_without_a_vehicle	18.68	15.29	**14.97**	18.58	19.46	22.13	25.5
Working_on_a_metal_crafts_project	26.89	22.56	**18.46**	20.79	28.02	33.98	31.28
Mean	31.23	28.28	**26.98**	28.60	30.76	33.93	36.83

By employing motion relativity, RMH-BoW performs significantly better than MH-BoW for most events. On one hand, compared with MH-BoW, RMH-BoW is invariant to the camera movement. On the other hand, RMH-BoW encodes not only the motion of objects, but also the interactions between different objects. This shows to be important in representing and recognizing different event occurrences. The experiment results demonstrate the effectiveness of our approach in capturing the object interactions in event detection compared with other motion features. By expanding visual words to the related words, ERMH-BoW slightly but consistently improves the performance of event detection. The word expansion alleviates the ambiguity caused by SIFT quantization where the same objects in an event with similar visual appearance would be mapped to different visual words. Table II also compares the performances of two representation approaches for the ERMH-BoW feature described in Section V. Our experiments show that EMD kernel performs better than RBF kernel. In the ERMH-BoW feature, we only encode the motion between neighboring keyframes. To some extent, this ignores the sequence information in event evolution. By matching the sequences of motion histograms in video events, EMD kernel compensates the weakness of ERMH-BoW in representing the sequence information and thus achieves better performance.

A.2 Performance of Feature Selection for ERMH-BoW: In this section, we present the performances of our approach described in Section IV for reducing the dimensionality of the ERMH-BoW feature. Fig. 5 shows the mean P_{MD} values for 20 events when $P_{FA} = 5\%$ and the total computation time for SVM training and classification when different thresholds are used to remove the less useful motion words. In Fig. 5, information gain is employed to remove the motion words which are weak in discriminating different event categories. As can be seen in Fig. 5, when 80% of the words are removed, the P_{MD} value is not or very slightly changed. This shows that most of the words are actually useless or even noisy, and can be treated as stop words for event detection. Meanwhile, with more motion words being removed, the computation time is reduced which is basically linear to the number of motion words. In our implementation, we empirically select 80% as the threshold to remove the less useful words in the information gain approach.

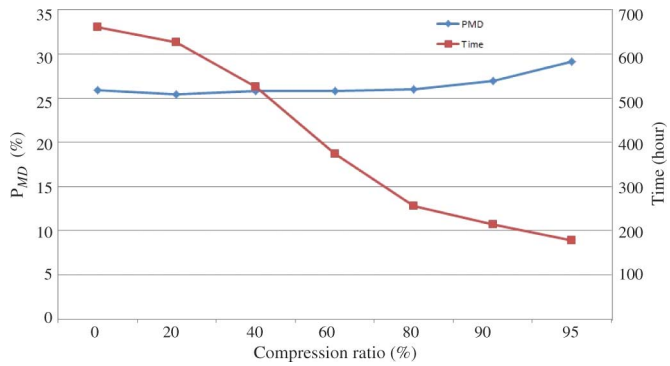


Fig. 5. Detection accuracy (mean p_{MD} values when $p_{FA=5\%} = 5\%$) and computation time when different thresholds are used for feature selection with information gain.

In the remaining motion words, we further employ the approach described in Section IV-B to measure their informativeness for the detection of each specific event. We assign different weights to different words and remove those words with weights smaller than a threshold. Fig. 6 shows the P_{MD} values and computation time when different weight thresholds are used. As can be observed in Fig. 6, when the less informative words are removed, the detection accuracy is slightly improved (P_{MD} value is reduced). This is because: i) The removed words are less useful in detecting the given events; ii) By assigning larger weights to the most informative words, the discriminative ability of the SVM is improved. However, the detection accuracy is reduced when too many words or some useful ones are removed. Similar to the information gain approach, the computation time is reduced when less words are used. In our implementation, we select 0.5 as the threshold in this approach to remove the less useful words.

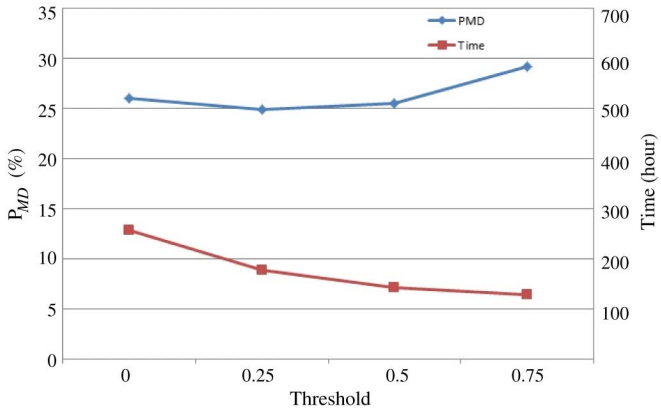


Fig. 6. Detection accuracy (mean p_{MD} values when $p_{FA=5\%}$) and computation time when different thresholds are used for feature selection with kernel optimization.

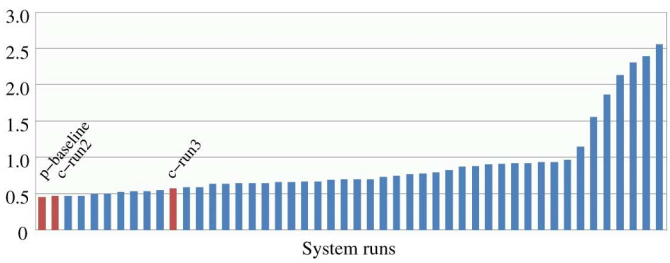


Fig. 7. Performance of our runs submitted to MED12 task. To better compare different systems, the last run with the highest NDC (9.30) is not shown in this figure.

B. On MED12 Progress Test Collections

In this section, we present the evaluation results during our participation in TRECVID 2012 Multimedia Event Detection task. In this evaluation, the Progress Test collections were used. We submitted three runs for the official evaluation [45]. Fig. 7 shows the performances of our submitted runs (in red color) among all submissions in term of actual NDC (Normalized Detection Cost) metric [56], [57]. The run c-run3 employs mainly the static visual, audio and semantic information in the videos. Visual features including SIFT and ColorSIFT are extracted in the sampled keyframes. For audio feature, MFCC coefficients are extracted in every audio frame of 50 ms, where each frame overlaps with its neighbors by 25 ms. 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 for three features respectively. Soft weighting is used for word assignment by mapping each descriptor to the 3 nearest words. Besides low-level features, we extract mid-level features, i.e. concept scores for event detection. Due to the lack of annotations on MED development set, we simply borrow the concept detectors constructed in the Semantic Indexing task [45]. In total, 46 concepts which are semantically related to the events are manually selected. The scores output by the corresponding detectors are used as the semantic features. For classifier learning, LIBSVM [7] is employed. Linear weighted fusion is used to combine all SVM outputs. In the submission, the thresholds are determined by minimizing the NDC scores.

TABLE III
NUMBER OF EVENTS MEETING DEFINED GOALS. THE GOALS ARE DEFINED TO BE MET IF THE SYSTEM'S $P_{Miss} < 4\%$ AND $P_{FA} < 50\%$

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

In the run c-run2, ERMH-BoW feature described in Section III is employed for event detection. The results are combined with c-run3 by late fusion. According to the evaluation results, the actual NDC value is reduced from 0.57 to 0.47. This shows the effectiveness of ERMH-BoW for detecting video events, especially for those events with intensive motion such as *bike_trick*, *parkour*, *winning_a_race*, and *grooming_animal*.

In the run p-baseline, feature selection is further employed to reduce the dimensionality of the ERMH-BoW feature by eliminating the less useful features. According to the evaluation results, the performance is a little bit improved by reducing the actual NDC from 0.47 to 0.45. 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 cleaner set of features, larger weights are assigned to the most important features and the performance is thus improved.

Among all the submitted runs, our two runs employing the ERMH-BoW feature (p-baseline and c-run2) are ranked 1 and 2 respectively. In other systems, most existing features including SIFT, ColorSIFT, Gist, MoSIFT, and DTF-HoG are employed. Different machine learning strategies, experimental settings, and their combinations are investigated [1], [6], [9], [34], [37], [50]. Although the detailed results and comparisons are not currently available due to the evaluation strategies in the MED task, the evaluation results have demonstrated the effectiveness of our approach based on motion relativity and feature selection for event detection compared with the state-of-the-art approaches.

Table III shows the number of events for which our systems meet the defined goals (the goals are met if the system's $P_{MD} < 4\%$ and $P_{FA} < 50\%$) with two methods for selecting the detection threshold in Equations (18) and (19). The thresholds for *Actual Decision* are selected on the development set by minimizing the NDC values, while the thresholds for *Target Error Ratio* are selected during the NIST evaluation at the intersection points between the system's DET curves and the Target Error Ratio lines [57]. As can be seen in Table III, our systems meet the defined goals for most of the twenty events. This shows that our approach can be successfully applied to the detection of various events.

VII. CONCLUSION

In this paper, we address event detection in open video domains by proposing a new motion feature namely ERMH-BoW. Bag-of-visual-words are adopted to represent *what* aspect of an event, while the relative motion histograms between visual words are used to capture the object activities or *how* aspect of the event. The derived feature ERMH-BoW can thus provide

a complete description of an event by closely integrating *what* and *how* aspects. By employing motion relativity, ERMH-BoW is invariant to varying camera movement. Furthermore, it captures not only the activity of each isolated object, but also the interactions between different objects/scenes to describe the event occurrences. Our experiments show that ERMH-BoW can significantly improve the performance of video event detection compared with the state-of-the-art approaches. To address the high-dimensionality problem with ERMH-BoW, we propose an approach for feature selection by measuring the discriminative ability of different motion words and removing the less useful ones. By feature selection, the computation time for event detection with the ERMH-BoW feature is reduced while the accuracy is also slightly improved. Similar to the traditional Bag-of-Visual-Words approach for image annotation, the spatial information is not considered in our approach. This would be our future direction so as to better represent the objects and their activities in videos. Furthermore, other models besides motion histograms will be investigated to capture the dynamic information in videos when employing the motion relativity for effective and efficient event detection.

REFERENCES

[1] R. Aly, K. McGuinness, and S. Chen *et al.*, "AXES at TRECVid 2012: KIS, INS, and MED," in *Proc. NIST TRECVID Workshop*, 2012.

[2] E. Ardizzone, M. L. Cascia, A. Avanzato, and A. Bruna, "Video indexing using MPEG compensation vectors," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, 1999, vol. 2.

[3] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[4] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Computer Vision*, 2006.

[5] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *Proc. Int. Conf. Computer Vision*, 2005.

[6] L. Cao, S. F. Chang, and N. Codella *et al.*, "IBM research and Columbia University TRECVID 2012 multimedia event detection (MED), multimedia event recounting (MER), and semantic indexing (SIN) systems," in *Proc. NIST TRECVID Workshop*, 2012.

[7] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," 2001 [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm.

[8] M. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," in *CMU-CS-09-161*, 2009.

[9] H. Cheng, J. Liu, and S. Ali *et al.*, "SRI-sarnoff aURORA aystem at TRECVID 2012: Multimedia event detection and recounting," in *Proc. NIST TRECVID Workshop*, 2012.

[10] N. Cristianini, J. Kandola, A. Elisseeff, and J. S-Taylor, "On kernel target alignment," *Adv. Neural Inf. Process. Syst.*, vol. 14, pp. 367–373, 2002.

[11] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Computer Vision*, 2006.

[12] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *Proc. IEEE Workshop Detection and Recognition of Events in Video*, 2001.

[13] Y. F. Day, S. Dagtas, M. Iino, A. Khokhar, and A. Ghafoor, "Object-oriented conceptual modeling of video data," in *Proc. Int. Conf. Data Engineering*, 1995.

[14] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2006.

[15] J. Gemert, C. Veenman, A. Smeulders, and J. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283.

[16] J. M. Geusebroek and G. J. Burghouts, "Performance evaluation of local colour invariants," *Comput. Vision Image Understand.*, pp. 48–62, 2009.

[17] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[18] A. Haubold and M. Naphade, "Classification of video events using 4-dimensional time-compressed motion features," in *Proc. Int. Conf. Image and Video Retrieval*, 2007.

[19] Y. G. Jiang, C. W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. Int. Conf. Image and Video Retrieval*, 2007.

[20] Y. Jiang, Q. Dai, X. Xue, W. Liu, and C. W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *Proc. Eur. Conf. Computer Vision*, 2012.

[21] Y. G. Jiang and C. W. Ngo, "Bag-of-visual-words expansion using visual relatedness for video indexing," in *Proc. ACM SIGIR*, 2008.

[22] Y. G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S. F. Chang, "Columbia-UCF TRECVID 2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in *Proc. NIST TRECVID Workshop*, 2010.

[23] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. Int. Conf. Computer Vision*, 2005.

[24] K. Kesorn and S. Poslad, "An enhanced bag-of-visual-word vector space model to represent visual content in athletics images," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 211–222, 2012.

[25] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2008.

[26] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[27] Z. Li, Y. Fu, T. S. Huang, and S. Yan, "Real-time human action recognition by luminance field trajectory analysis," in *Proc. ACM Multimedia Conf.*, 2008, pp. 671–675.

[28] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, 2004.

[29] B. D. Lucas and T. Kanade, "An interactive image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artifical Intelligence*, 1981, pp. 121–130.

[30] Y. F. Ma and H. J. Zhang, "Motion pattern-based video classification and retrieval," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 1, pp. 199–208, 2003.

[31] L. Zelnik-Manor and M. Irani, "Statistical analysis of dynamic actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1530–1535, 2006.

[32] A. F. Martin, G. Doggington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997.

[33] K. Mikoljczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vision*, vol. 60, pp. 63–86, 2004.

[34] M. Murata, T. Izumitani, and H. Nagano *et al.*, "NTT communication science laboratories and national institute of informatics at TRECVID 2012: Instance search and multimedia event detection tasks," in *Proc. NIST TRECVID Workshop*, 2012.

[35] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[36] P. Natarajan and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[37] P. Natarajan, P. Natarajan, S. Wu, and X. Zhuang *et al.*, "BBN VISER TRECVID 2012 multimedia event detection and multimedia event recounting systems," in *Proc. NIST TRECVID Workshop*, 2012.

[38] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[39] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[40] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2009.

[41] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

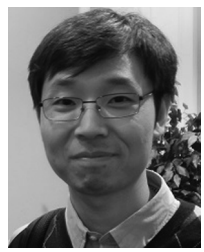
[42] A. Torralba and A. Oliva, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[43] F. Wang and B. Merialdo, "Weighting informativeness of bag-of-visual-words by kernel optimization for video concept detection," in *Proc. Int. Workshop Very-Large-Scale Multimedia Corpus, Mining and Retrieval*, 2010.

[44] F. Wang, Y. Jiang, and C. W. Ngo, "Video event detection using visual relatedness and motion relativity," in *Proc. ACM Multimedia Conf.*, 2008.
[45] F. Wang, Z. Sun, D. Zhang, and C. W. Ngo, "Semantic indexing and multimedia event detection: ECNU at TRECVID 2012," in *Proc. NIST TRECVID Workshop*, 2012.
[46] H. Wang, A. Klser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2011.
[47] X. Wu, C. W. Ngo, J. Li, and Y. Zhang, "Localizing volumetric motion for action recognition in realistic videos," in *Proc. ACM Multimedia Conf.*, 2009.
[48] D. Xu and S. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1985–1997, 2008.
[49] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. Int. Conf. Machine Learning*, 1997.
[50] S. Yu, Z. Xu, D. Ding, and W. Sze *et al.*, "Informedia@TRECVID 2012," in *Proc. NIST TRECVID Workshop*, 2012.
[51] T. Zhang, C. Xu, G. Zhu, S. Liu, and H. Lu, "A generic framework for event detection in various video domains," in *Proc. ACM Multimedia Conf.*, 2010.
[52] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong, "Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor," in *Proc. ACM Multimedia Conf.*, 2009.
[53] K. Li, J. Hu, and Y. Fu, "Modeling complex temporal composition of actionlets for activity prediction," in *Proc. Eur. Conf. Computer Vision*, 2012.
[54] I. Everts, J. C. Gemert, and T. Gevers, "Evaluation of color STIPs for human action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.
[55] TRECVID [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid.
[56] TRECVID, "Multimedia event detection track," [Online]. Available: http://nist.gov/itl/iad/mig/med.cfm
[57] J. Fiscus, "2012 TRECVID multimedia event detection evaluation plan," [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/MED12-EvalPlan-V03.pdf.

Feng Wang received his PhD in Computer Science from the Hong Kong University of Science and Technology in 2007 and BsC from Fudan University, China, in 2001 respectively. Before joining East China Normal University as an associate professor in the Dept. of Computer Science and Technology, he was a research fellow in City University of Hong Kong and Institute Eurecom, France. His research interests include multimedia information retrieval, pattern recognition and IT in education.

Zhanhu Sun received his BSc degree from Shanxi University, Shanxi Province, China in 2010. He is currently a Master student in the Dept. of Computer Science and Technology, East China Normal University, Shanghai, China. His research topic is video event detection and multimedia information retrieval.



Yu-Gang Jiang received the Ph.D. degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2009. During 2008-2011, he was with the Department of Electrical Engineering, Columbia University, New York. He is currently an Associate Professor of computer science with Fudan University, Shanghai, China. His research interests include multimedia retrieval and computer vision. Dr. Jiang is an active participant of the Annual U.S. NIST TRECVID Evaluation and has designed a few top-performing video analytic systems over the years. He is also an organizer of the THUMOS action recognition challenge and the Violent Scenes Detection Task of the MediaEval benchmark. His work has led to several awards including the 2013 ACM Shanghai Distinguished Young Scientist Award.

Chong-Wah Ngo (M'02) received the B.Sc. and M.Sc. degrees in computer engineering from Nanyang Technological University, Singapore, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong. He was a Post-Doctoral Scholar with the Beckman Institute, University of Illinois at Urbana-Champaign, Champaign. He was a Visiting Researcher with Microsoft Research Asia. He is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. His current research interests include large-scale multimedia information retrieval, video computing and multimedia mining.

Dr. Ngo is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA. He is the Conference Co-Chair of the ACM International Conference on Multimedia Retrieval 2015 and Pacific Rim Conference on Multimedia 2014; and the Program Co-Chair of the ACM Multimedia Modeling Conference 2012 and the ACM International Conference on Multimedia Retrieval 2012; and the Area Chair of the ACM Multimedia 2012. He was the Chairman of the ACM (Hong Kong Chapter) from 2008 to 2009.