Video Event Detection Using Motion Relativity and Visual Relatedness

Feng Wang Yu-Gang Jiang Chong-Wah Ngo
Dept. of Computer Science, City University of Hong Kong
83 Tat Chee Avenue, Kowloon Tong, Hong Kong
{fwang, yjiang, cwngo}@cs.cityu.edu.hk

ABSTRACT

Event detection plays an essential role in video content analysis. However, the existing features are still weak in event detection because: i) most features just capture *what* is involved in an event or *how* the event evolves separately, and thus cannot completely describe the event; ii) to capture event evolution information, only motion distribution over the whole frame is used which proves to be noisy in unconstrained videos; iii) the estimated object motion is usually distorted by camera movement. To cope with these problems, in this paper, we propose a new motion feature, namely Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW) to employ motion relativity and visual relatedness 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 visual words to depict the object activities or *how* aspect of the event. ERMH-BoW thus integrates both *what* and *how* aspects for a complete event description. Instead of motion distribution features, local motion of visual words is employed which is more discriminative in event detection. Meanwhile, we show that by employing relative motion, ERMH-BoW is able to honestly describe object activities in an event regardless of varying camera movement. Besides, to alleviate the visual word correlation problem in BoW, we propose a novel method to expand the relative motion histogram. The expansion is achieved by diffusing the relative motion among correlated visual words measured by visual relatedness. To validate the effectiveness of the proposed feature, ERMH-BoW is used to measure video clip similarity with Earth Mover's Distance (EMD) for event detection. We conduct experiments for detecting LSCOM events in TRECVID 2005 video corpus, and performance is improved by 74% and 24% compared with existing motion distribution feature and BoW feature respectively.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

(a) (b) (c)

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

General Terms

Algorithm, Experimentation, Performance.

Keywords

Video event detection, motion relativity, visual relatedness.

1. INTRODUCTION

With the multimedia information widely available from different sources such as web, the management and retrieval of multimedia data has been actively researched in the past few decades, where the 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 [2, 4, 34, 35], action classification [6, 10, 11, 18, 23, 30, 32], and event recognition [9, 12, 17, 31].

Recently semantic detection has attracted a lot of attentions. In the high-level feature extraction task of annual TRECVID workshop [37], 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 certain success has been achieved, the result is still far away from satisfactory due to the bottleneck of large gap between semantic and low-level features. On the other hand, the event-based concepts specifically have not been paid enough attentions and the performance is still poor.

In contrast to static concepts, event has its own nature, i.e. dynamic nature. As a result, besides the semantic gap which exists in the detection of all concepts, event-based concept detection is also limited by the keyframe-based approaches that are widely used for static concept detection. Without viewing the dynamic course of the event, human frequently

Figure 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.

encounter difficulties in event annotation. In [36], by comparing the two-round manual annotations of 24 events based on keyframe and video sequence respectively, only about 78% of the keyframe based annotations are correct. For those motion-intensive events, the accuracy is even lower such as *Dancing* (42%) and *People_Marching* (39%). Figure 1 shows the difficulty in keyframe-based event annotation and detection. For instance, in (a), by looking at the keyframe only, even for 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 becomes necessary to employ sequence information in event-based concept detection instead of the keyframe only.

In this paper, we focus on extracting effective features from video sequence for event detection. In a video clip, an event is usually described from two aspects: i) *what* are involved in the event, *e.g.* person, objects, buildings, etc; ii) *how* the event evolves in temporal domain, i.e. the course of the event. The former consists of 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 (bag-of-visual-words), 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 objects and the interaction among different people. This information can only be captured by viewing the whole frame sequence. Motion is an important cue in describing event evolution. Recently, various motion features have been developed to capture motion information in the sequence such as motion histogram [6] and motion vector map [12]. However, the existing features are still weak in event-based concept detection because: i) most features only consider one of the two aspects, i.e. exploit *what* or *how* separately and thus cannot completely describe an event; ii) only motion distribution information is used, which has been proven to be noisy in unconstrained videos; iii) the observed motion in the video clip is distorted by camera movement, and cannot depict the real object activities and interactions in an event.

Figure 2 shows two example 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 static features (color moment, bag-of-visual-words) and sequence feature (motion histogram) respectively. However, neither one of them is enough to describe event *Walking*. Furthermore, motion is usually distorted by varying camera movement. For instance, in the second clip of Figure 2, camera follows the person when he walks through the yard. No motion of person can be detected by traditional motion estimation. Therefore, motion calculated in frame sequence cannot honestly present the real activities of the person. However, in both two clips, we can see that the relative position between *Person* and background scene (*Building*) is changing, and similar relative motion patterns can be observed under various camera movements. Thus, relative motion is suitable to cope with camera movements for event description and detection.

Due to its ability to honestly describe 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). Figure 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) which has been proven to be effective in concept detection [3, 13], 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, given a video frame, it can be represented as a visual word histogram by mapping its keypoints to the visual vocabulary. With BoW, the activities and interactions between objects can be captured by modelling the relative motion between visual words depicting different objects/scenes. For instance, as illustrated in Figure 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 event *Walking*. As shown

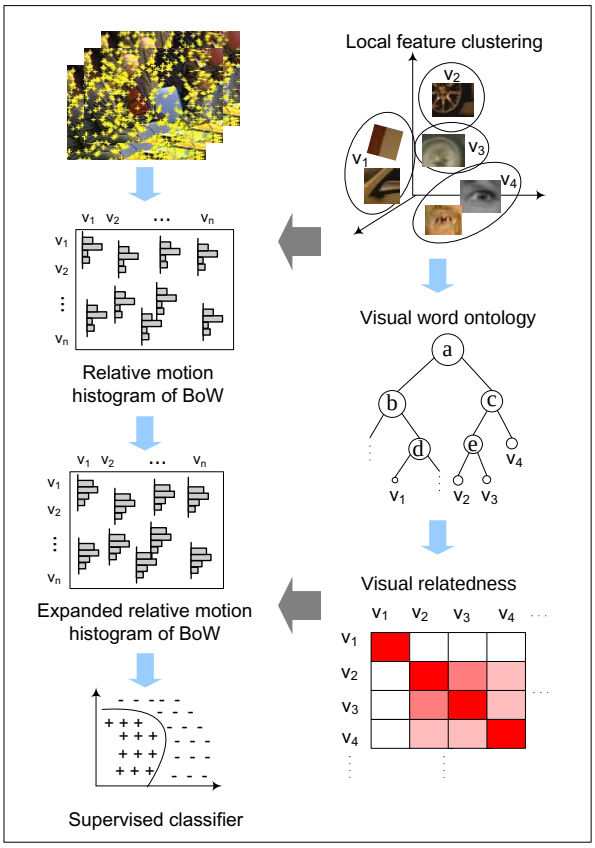

Figure 3: Proposed feature extraction framework for video event detection. With visual words capturing what are involved in an event, the local motion histogram of visual words describes both *what* and *how* aspects effectively for a complete representation of an event. Motion relativity and visual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

on the left of Figure 3, given a video clip, the keypoints are tracked in neighboring frames and relative motion is calculated between every two keypoints. Given two visual words, a relative motion histogram (RMH-BoW) is constructed by accumulating motion vectors between every two keypoints mapped to the two words respectively (Section 3).

However, with motion relativity between visual words representing activities between objects/scenes, a feature mismatch problem may be caused. As seen in Figure 2, although both keypoints Q_{p1} and Q_{p2} actually depict *Person*, they cannot be matched when they are mapped to different visual words in BoW representation. This is possible considering the fact that visual words are the outcome of clustering algorithm, and can be correlated to each other due to the quantization effect. In this case, although they show similar motion patterns of the same object category *Person*, activities of different objects are detected. To alleviate this visual word correlation problem, as shown on the right of Figure 3, we first construct a visual word ontology to measure the relatedness between different words [14]. The visual word relatedness is then incorporated into RMH-BoW to derive a new feature called Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW; detailed in Section 4). The expansion of RMH-BoW is achieved by diffusing the relative motion between two visual words to other words that are correlated to them, i.e. with higher visual relatedness. Finally, ERMH-BoW is used for supervised learning and event detection (Section 5).

Compared with existing features, the novelty of ERMH-BoW lies in the following aspects. First, the two aspects of an event, i.e. *what* and *how*, are closely integrated to completely describe events. Second, local motion is used, which is more discriminative in event detection. Third, motion relativity is exploited to honestly depict the activities in an event. Finally, visual word relatedness is employed to expand the relative motion histogram, which can effectively alleviate the correlation problem in the widely used BoW feature.

2. RELATED WORKS

In [9], 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. This approach achieves some improvement compared with keyframe based approach from an experiment on a small set of events. In [31], 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 video similarity. EMD distance is then incorporated into the kernel function of SVM framework for event detection. While these approaches attempt to construct different models for describing event evolution, in this paper, we focus on extracting effective features, specifically motion features for event detection.

As an important cue to characterize video content, motion analysis has been intensively studied for video indexing and retrieval. In [1], 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 [24], a motion pattern descriptor namely motion texture is proposed for video retrieval and the classification of simple camera and object motion patterns. In [8], spatio-temporal interactions between objects are expressed by predicate logic for video retrieval. This algorithm assumes the objects are correctly detected and located during video preprocessing. However, automatic object segmentation and semantic annotation remains extremely difficult and unreliable in unconstrained videos.

Many events can be represented as 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, and has been employed in some previous works in order to capture the event evolution information. In [7], a ground-based mobile surveillance system is built to detect and track moving people. Their activities are recognized by PCA-based matching. In [6], Motion History Image (MHI) is calculated over a frame sequence to describe the characteristic of human motion. Recognition is achieved by statically matching MHIs. This approach is applied to well-segmented human figures for recognizing several predefined actions. In [17], event is treated

as a space-time volume in the video sequence. Volumetric features based on optical flow are extracted for event detection. The approach is used in videos with single moving object (human) and action. In [29], a similarity measure is proposed to search for two different video segments with similar motion fields and behaviors, while recognition is not performed. Although some success has been achieved, these algorithms are just employed in some specific video domains (such as surveillance videos) in known environments or simple scenarios with only the object of interest. The motion distribution features could be easily noised by those motion of no interest if being applied to general videos with different moving objects.

Some attempts have also been made to employ motion for event recognition in unconstrained videos. In [12], 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. However, more events with different motion intensities and patterns should be investigated on a larger video corpus. In [5], motion is employed in searching for some event-based topics, for instance, "*Find shots of one or more people walking up stairs*". Eight directions of motion vectors and intensities are efficiently extracted from motion vectors in MPEG compressed domain and exploited for final ranking. The MAP for 24 topics is improved from 0.04 to 0.043 by combining motion feature with text, concept and visual features. Due to the overall low accuracy in search task, it is difficult to draw the conclusion whether or by how much the results could be indeed improved.

In summary, the problems with these existing features lie in three aspects. First, motion is separated from the static visual information, or *what* aspect of an event (such as car, person) and considered independently. As a result, motion might be useful in discriminating those events with different motion patterns. However, different events may show similar motion of different objects. For instance, *Person_Running* and *Car_Running* could have the same motion patterns in horizontal direction. If only motion is used without considering what is running, they will be detected as the same event. To fully describe an event, the two aspects should be closely integrated. Second, motion feature is extracted to describe the motion distribution in the whole frame. This might work in some specific video domains such as surveillance videos in a lift, or videos containing few objects, where the objects of interest dominate the motion in the video. In an unconstrained videos, the motion distribution could be greatly noised, and weak in event description. To effectively capture event evolution information, local motion specifically the motion of event-centered objects should be highlighted. Third, due to the distortion from camera movement, the motion vectors calculated from video sequence is composed of object motion plus camera movement. Thus, it cannot depict the real object activities. This problem has not been carefully studied.

In [19], inspired by the success of bag-of-visual-words (BoW) in concept detection, space-time interest points with salient changes in both spatial and temporal domains are extracted in videos. In [27], spatial-temporal words are then derived for recognizing human actions. This feature somehow addresses the first two problems by employing local motion and capturing the information in both spatial and temporal domains. However, experiments are just conducted on simple video sequences captured by stationary camera to recognize only few actions of human. As we know, spatial-temporal word has not been successfully employed in unconstrained videos. One important reason is that the space-time interest point is not invariant to camera movement. In this paper, we show that relative motion is able to completely and honestly describe the events compared with these existing features. By employing motion relativity, the proposed feature ERMH-BoW can cope with all the problems mentioned above for event description and significantly improve the performance of detection.

3. RELATIVE MOTION HISTOGRAM OF BoW (RMH-BoW)

As discussed in Sections 1, relative motion between different objects and scenes is able to honestly describe an event compared with existing features. In this section, we propose to employ motion relativity for developing effective features in event detection. We first briefly describe the generation of BoW, which is used to capture *what* aspect. Given a video clip, we then construct a motion histogram for each visual word (MH-BoW) by employing local motion information. Finally, we modify MH-BoW by replacing the motion vectors with the relative motion between different visual words so as to capture motion relativity.

3.1 BoW Generation

Keypoints are salient patches containing rich local information of images. Shown as small crosses in the images on the left of Figure 3, they usually lie around the corners and edges in image objects. In this paper, we use DoG [25] as keypoint detector and SIFT [21] as keypoint descriptor.

In BoW image representation, firstly a visual vocabulary is generated by using k-means algorithm to cluster a set of keypoints (SIFT features) and each cluster is treated as a visual word. By mapping the keypoints in an image to the vocabulary [13], we can represent the image as a vector of visual words with its weights indicating the presence or absence of the visual words.

3.2 Motion Histogram of Visual Words (MH-BoW)

Instead of motion distribution which could be easily noised in unconstrained videos, we employ the local motion information to derive more discriminative features. In this section, we construct a local motion histogram for each visual word in BoW. Given a video clip, our motion features are extracted between every two neighboring frames. To be efficient, 5 frames are evenly sampled every second. For a sampled frame at time t, keypoints are first detected by DoG [25]. We then employ the algorithm in [22] to track the keypoints in the next frame. For each keypoint p that can be successfully tracked, we calculate its motion vector m_p 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_p is decomposed into four components

$D_i(m_p)$, where $i = 1, 2, 3, 4$ are corresponding to the four directions: left, right, up and down; and $D_i(.)$ projects m_p to the i^{th} direction. For a visual word v, the motion histogram is calculated as

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

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

By Equation 1, we get an N-dimension feature vector called Motion Histogram of BoW (MH-BoW), where N is the number of visual words, 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. Thus, by MH-BoW, we have addressed two problems that lie in the current existing features, i.e. how to integrate *what* and *how* aspects of an event, and how to make use of more discriminative local motion information for event description.

3.3 Relative Motion Histogram between Visual Words (RMH-BoW)

Based on RM-BoW, in this section, we address the third problem in feature extraction, i.e. the motion distortion problem caused by varying camera movement. To this end, we propose to employ motion relativity between different objects and scenes. As discussed in Section 1 and observed in Figure 2, motion relativity can cope with the motion distortion problem caused by camera movement, and remain consistent for different clips containing the same event regardless of different camera movement. In other words, it is able to honestly describe the real activities in an event. With BoW representing *what* aspect of an event, we capture the object activities by using the relative motion between different visual words. We modify MH-BoW in Section 3.2 with the relative motion histograms between visual words.

Given two visual words v_a and v_b, the relative motion histogram between them is calculated as

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

where p and q are two interest points mapped to visual words v_a and v_b respectively, $m_p - m_q$ is the relative motion of p with reference to q, and $D_i(.), i = 1, 2, 3, 4$ decomposes the relative motion vector to the four directions as in the last section to generate a 4-directional histogram between visual words v_a and v_b. By Equation 2, the motion information in a video clip is represented as an $N \times N$ matrix $\mathcal{R}$, where each element $\mathcal{R}(i, j)$ is a relative motion histogram between the two visual words i and j. 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.

4. EXPANDING RMH-BoW WITH VISUAL WORD RELATEDNESS (ERMH-BoW)

As discussed in Section 1, when BoW is used to represent the *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. 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 propose a method to expand the relative motion histogram based on visual relatedness. The expansion is conducted by diffusing the motion histograms across correlated visual words, so that the visual word correlation problem can be alleviated.

4.1 Visual Relatedness

The visual relatedness is a measurement of visual word similarity, which can be estimated in the same way as we estimate the semantic relatedness of textual words using general ontology such as WordNet. We apply the method of [14] to estimate the visual relatedness. Given a set of keypoints, we first construct a visual vocabulary through clustering the keypoints by k-means algorithm. With the visual vocabulary, a visual ontology is further generated by adopting agglomerative clustering to hierarchically group two 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 Figure 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 [16], to estimate the visual relatedness. Denote v_i and v_j as two visual words, JCN considers the ICs of their common ancestor and the two compared words, defined as:

$$\text{JCN}(v_i, v_j) = \frac{1}{\text{IC}(v_i) + \text{IC}(v_j) - 2 \cdot \text{IC}(\text{LCA}(v_i, v_j))} \tag{3}$$

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

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

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

Compared to directly calculating the visual word relatedness based on the proximity of cluster centers, using JCN can be more accurate as not only cluster proximity, but also the cluster size/density (inferred by IC) are taken into account.

4.2 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

$$\mathcal{R}^E(i,j) = \mathcal{R}(i,j) + \sum_{s_i,s_j} \text{JCN}(s_i,i) \times \mathcal{R}(s_i,s_j) \times \text{JCN}(s_j,j) \tag{5}$$

where $\{s_i\}$ and $\{s_j\}$ are the sets of visual words that are correlated to the words i and j 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 Figure 2, if the two points Q_{p1} and Q_{P2} lying on *Person* were assigned to different visual words, say v_1 and v_2 respectively, which 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 and also retains the sparse property of the resulted ERMH-BoW histograms. On the other hand, diffusing with more visual words does not promise better performance.

5. EVENT DETECTION

In this section, we employ the proposed feature ERMH-BoW for event detection to demonstrate the effect of motion relativity and visual relatedness. For this purpose, we adopt the kernel based algorithm in [30]. First, ERMH-BoW is used to measure the distance between different video clips. This distance is then incorporated into the kernel function of support vector machine (SVM) for event detection.

5.1 Distance between Video Clips

First, given a video clip, ERMH-BoW is calculated between every two neighboring sampled frames. The video clip is then represented by a sequence of ERMH-BoW as its signature. For video clip similarity measure, EMD (Earth Mover's Distance) has proven to be effective in clip alignment and matching [28]. To employ EMD for video clip similarity measure, the ground distance between a pair of frames from two clips is defined as the Euclidean distance of the ERMH-BoW corresponding to the two frames:

$$d(\mathcal{R}_1^E, \mathcal{R}_2^E) = \sqrt{\frac{1}{N^2} \sum_{1 \le i,j \le N} (\mathcal{R}_1^E(i,j) - \mathcal{R}_2^E(i,j))^2} \tag{6}$$

Given two clips $A = \{(\mathcal{R}_{A1}^E, w_{A1}), \cdots, (\mathcal{R}_{Am}^E, w_{Am})\}$ and $B = \{(\mathcal{R}_{B1}^E, w_{B2}), \cdots, (\mathcal{R}_{Bn}^E, w_{Bn})\}$ with m and n ERMH-BoW matrices as signatures respectively, and $w_{Ai} = 1/m$, $w_{Bj} = 1/n$ as the weights for each ERMH-BoW matrix. The EMD distance between A and B is computed by

$$D(A,B) = \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} d(\mathcal{R}_{Ai}^E, \mathcal{R}_{Bj}^E)}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \tag{7}$$

where f_{ij} is the optimal match among two sequences of ERMH-BoW matrices of A and B. The details of how to determine f_{ij} can be found in [31].

Considering the large scale of the motion histogram, the distance measure could be quite slow. Fortunately, this problem can be alleviated by employing the sparseness property of the feature matrix. Furthermore, each clip usually contains at most tens of sampled frames and EMD matching can be carried out efficiently.

5.2 SVM Detection

With the EMD distance between video clips computed in Equation 7 by employing ERMH-BoW features, we adopt the algorithm in [31] to train SVMs for event detection. The EMD distance between video clips is incorporated into the kernel function of the SVM framework by using Gaussian function:

$$K(A,B) = exp(-\frac{1}{\kappa M} D(A,B)) \tag{8}$$

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

In [31], 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 and compare the performance of ERMH-BoW with other existing features under the same framework for event detection.

6. EXPERIMENTS

In this section, we conduct video event detection experiments to validate the effectiveness of the proposed ERMH-BoW. Comparisons with other existing motion feature and static features will be given.

6.1 Data Description

We choose the LSCOM annotated events [36] in the experiments. This dataset is annotated specifically for event-based concept detection in news videos. The video data are from TRECVID 2005 benchmark, and 24 events are re-annotated by watching the video clip based on the preliminary keyframe-based annotation. This provides more accurate annotations since the presence of many events cannot be correctly judged by just looking at one or few static keyframes. After removing some events with only few positive examples, in our experiments, 14 events are used namely *Exiting_Car*, *Handshaking*, *Demonstration_Or_Protest*, *Riot*, *Running*, *Walking*, *Dancing*, *Shooting*, *Airplane_Flying*, *Election_Compaign_Greeting*, *Street_Battle*, ,*People_Marching*, *People_Crying*, and *Singing*. The definition of these events can be found in [36]. In the selected 14 events, the numbers of positive examples range from 61 to 2332. Negative examples for each event are composed of annotated negative clips and extra examples randomly selected from other event categories. As a result, there are totally about $40,000$ clips in the dataset, among which 50% is used for classifier training and the remaining for testing. For performance evaluation, we use Average Precision (AP) [37] which has been the official performance metric in TRECVID evaluation since 2001.

Table 1: Average Precision (%) for event detection with different features. CM: Color Moment; CS: Concept Score; GMH: Grid-based Motion Histogram; BoW: Bag-of-Visual-Words; MH-BoW: Motion Histogram of BoW; RMH-BoW: Relative Motion Histogram of BoW; ERMH-BoW: Expanded Relative Motion Histogram of BoW; MAP: Mean Average Precision.

Event	GMH	CM	CS	BoW	MH-BoW	RMH-BoW	ERMH-BoW
Exiting_Car	17.4	19.7	23.7	25.3	29.2	32.5	**34.8**
Handshaking	6.9	8.2	10.3	10.1	10.8	11.0	**11.4**
Running	32.8	48.6	57.1	56.7	61.2	66.7	**68.7**
Demonstration_Protest	27.5	29.2	31.3	31.6	33.4	34.7	**36.8**
Walking	20.7	24.3	29.8	28.5	32.4	37.6	**39.2**
Riot	10.3	13.7	16.5	18.7	20.6	22.5	**23**
People_Marching	21.4	20.8	23.4	24.3	26.9	28.9	**30.2**
Dancing	12.1	14.9	16.2	15.4	16.5	18.2	**19.5**
Election_Campaign_Greeting	7.4	9.1	10.4	10	11.3	12.9	**13.9**
Shooting	9.2	10.7	11.0	10.4	11.2	11.5	**12.4**
Airplane_Flying	13.6	14.1	17	17.3	19.1	21.6	**23.1**
Street_Battle	15.7	20.9	23.8	24.1	26.3	29.1	**32.3**
People_Crying	4.6	5.5	**8.6**	8.0	7.4	7.2	7.2
Singing	9.0	11.3	**12.1**	11.6	10.9	11.2	11.0
MAP	14.90	17.93	20.80	20.86	22.66	24.69	**25.96**

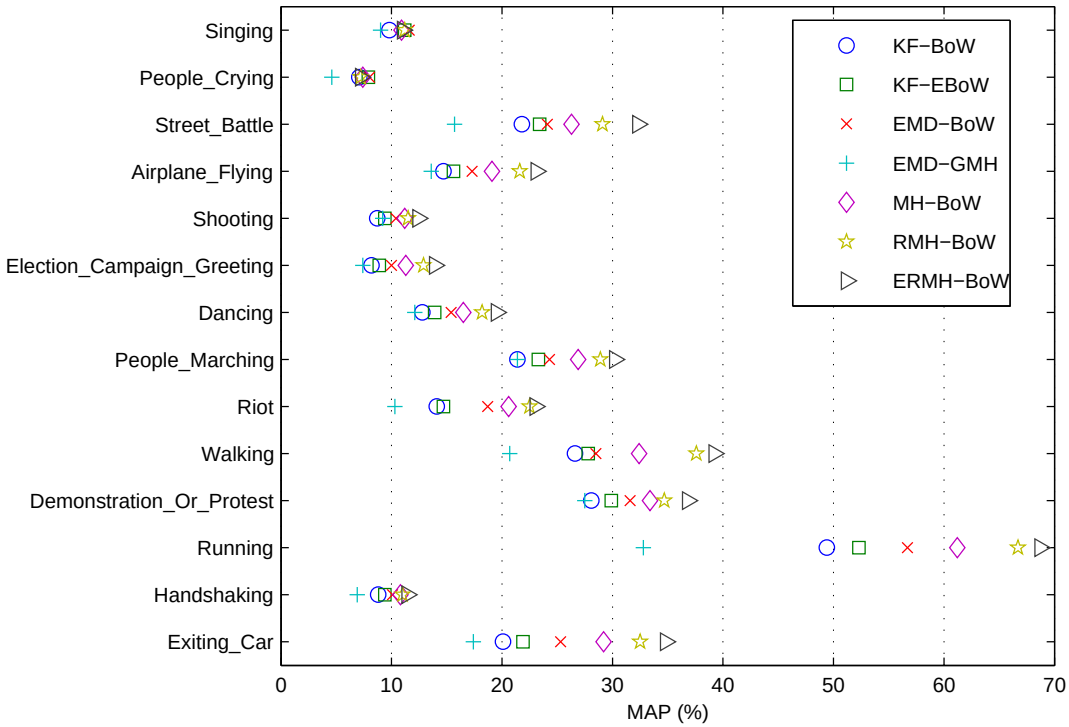

Figure 4: Per-event performance comparison of keyframe-based (KF) and sequence-based approaches.

6.2 ERMH-BoW vs. Motion Distribution

To compare ERMH-BoW with motion distribution features in event detection, we extract Grid-based Motion Histogram (GMH) from each frame in the sequence. Each frame is segmented into 5×5 grids in spatial domain. Motion vectors are extracted from MPEG compressed video and aggregated in each grid to form 4−directional histograms, which results in a 100-d feature vector. For event detection with GMH, the similar algorithm in Section 5 is employed by using GMH as the signatures of video clips.

As seen in Table 1, compared with GMH, ERMH-BoW improves the MAP (Mean Average Precision) of 14 events by 74.2% (from 14.90% to 25.96%). GMH just considers *how* aspect of an event, but ignores *what* aspect. Thus, it might be useful to discriminate those events with different motion intensities and patterns, *e.g. Singing* vs. *People_Marching*, but does not help much if different events have similar motion intensities and patterns. This is why current motion distribution features are just applied to some specific video domains or for recognizing a small set of events with different motion patterns. In ERMH-BoW, the presence of visual words captures *what* aspect of an event, while their corresponding motion histograms describe *how* aspect. Thus, ERMH-BoW encodes both two aspects of an event in a single feature to provide a complete description.

In addition, instead of motion distribution in GMH, ERMH-BoW takes the advantage of employing local motion information. Intuitively, only motion of event-centered object is useful for event description and detection. For instance, only the motion of a person can be used to detect an event *Walking*, while the movement of other objects could be random and noisy in unconstrained video domains. However, motion distribution feature simply sums up all detected motion in the spatial domain, which can be easily noised. Instead, ERMH-BoW employs the local motion of the visual words, and thus the activities of those objects of interest (depicted by visual words) can be highlighted during supervised learning. This makes ERMH-BoW a more discriminative feature in event detection. (How ERMH-BoW benefits from motion relativity will be discussed in the next section.)

6.3 ERMH-BoW vs. Static Features

We compare ERMH-BoW with the features that are widely used in concept detection, including Color Moment (CM), Bag-of-Visual-Word (BoW), and Concept Score (CS). For experiments, these features are extracted in each frame of the sequence used in ERMH-BoW extraction. CM is a global color feature. We calculate the first 3 moments of 3 channels in *Lab* color space over 5×5 grids for each keyframe, and aggregate the features into a 225-d feature vector. BoW makes use of local visual information in images. The generation of BoW is the same as in Section 3.1. The local patches (keypoints) are detected by DoG [25] and described with SIFT [21]. We build a vocabulary of 500 visual words by clustering the SIFT descriptors. TF-IDF is used to weight the significance of a word in the keyframe, resulting in a 500-d feature vectors. Besides these two low-level features, CS can be seen as a mid-level feature. Given a frame, the concept score vector is calculated by employing the VIREO-374 concept detectors [15]. In VIREO-374 detectors, three kinds of features: color moment, wavelet texture, and BoW (Bag-of-Words) are extracted for training three sets of classifiers respectively. Given a frame, the score for each concept is calculated as the mean of the three classifiers' outputs. We select 70 concept detectors which are related to the experimented events to compose a 70-d concept score vector. The extracted features are then used as the signatures of video clips as in Section 5 for EMD matching and SVM recognition.

As seen in Table 1, among these three features, CM gets the lowest MAP of 17.9%, while the performance of BoW (20.9%) and CS (20.8%) is similar. BoW outperforms CM by employing more discriminative local information in images. The result is consistent with concept detection [13]. CS employs BoW indirectly in concept score calculation, and probably this is why they achieve similar MAPs. Compared with BoW, ERMH-BoW improves the MAP by 24.4% (from 20.86% to 25.96%). These three static features just capture *what* aspect of an event, but ignore *how* aspect. Notice that EMD matching with static features (CM, BoW, CS) indeed models some sequence information of event evolution. However, by this experiment, we can see it is still necessary to develop features to describe the event evolution. This is because EMD matching just depicts event evolution coarsely without considering detailed event activities. For instance, when matching two clips using CS features, EMD just investigates the presence or absence of detected concepts along the sequence. Two clips containing the same objects will be regarded as similar although the object activities are totally different corresponding to different events.

To further evaluate the effect of motion relativity and visual relatedness, we also test the performance of another two features: MH-BoW and RMH-BoW extracted in Section 3.2 and 3.3 respectively. As seen in Table 1, the improvement of ERMH-BoW compared with BoW can be decomposed into three pieces due to local motion, motion relativity and visual relatedness respectively.

MH-BoW encodes local motion information of visual words. Compared with BoW, an improvement of 8.6% is achieved. On one hand, this shows the necessity of integrating *what* and *how* aspects for event detection. Motion, especially local motion, is useful in describing event evolution. On the other hand, by using MH-BoW, the improvement is not significant enough to convince the importance of motion considering the extra computation amount needed. This is because the motion vector calculated as spatial difference between frames in Section 3.2 cannot honestly depict the real object activities in an event. As shown in Figure 2, MH-BoW might work for the first clip, but not for the second one due to the varying camera movement. To cope with this problem, relative motion is used in RMH-BoW. With BoW representing *what* aspect in an event, object activies are encoded as the relative motion between different visual words. Compared with MH-BoW, another improvement of 9.0% is observed. This supports our argument in Section 1 that relative motion is better at honestly depicting object activities and interactions in event detection. The third piece of improvement of ERMH-BoW comes from the visual relatedness. Intuitively, in RMH-Bow, for two video clips containing the same event, some different visual words may belong to the same object or scene, but are treated independently. As a result, the two clips cannot be correctly matched as expected. In ERMH-BoW, we alleviate this problem by diffusing the relative motion histogram among correlated visual words based on visual relatedness. The experiment shows the effectiveness of

our approach. Compared with RMH-BoW, an improvement of 5.1% is achieved by ERMH-BoW.

As seen in Table 1, by employing the local motion and relative motion information, the improvement is mainly from those motion-intensive events such as *Running*, *Walking*, and *People_Marching*. Meanwhile, some decline is also observed for those events which are not motion-intensive, such as *Singing* and *People_Crying*. This is because there is no or very little motion involved, and motion information is less useful for detecting these events. In addition, the (relative) motion histograms of most visual words integrate visual and motion information together, and proves useful in capturing the motion of event-centered objects. However, for non-motion-intensive events, the motion for all visual words are almost equally minor, which cannot describe the event evolution and even weakens the effectiveness of BoW weighting scheme. As a result, the features with only static visual (BoW) or semantic (CS) information could perform better than the motion features for these two events.

6.4 Comparison to Keyframe-based Approaches

To give a more comprehensive understanding for the performance of the different features for event detection, we also conduct experiments to contrast sequence-based and keyframe-based (KF) event detection. For KF-based approaches, a single keyframe is used for feature extraction. The features used include CM, BoW, CS, and EBoW (Expaned BoW with Visual Relatedness). The first three features are the same as used in Section 6.3, but extracted just in the keyframe. EBoW is derived by diffusing the weight of each visual word in BoW using visual relatedness in a similar way as we expand RMH-BoW.

Table 2 shows the results of keyframe-based approaches. By comparing the results to Table 1, sequence based approaches perform much better than keyframe based approaches. Figure 4 plots the overall performances of sequence-based and KF-based approaches for the 14 tested events. Comparing the best performances of sequence-based (ERMH-BoW) and KF-based (EBoW) approaches, an improvement of nearly 35% is achieved when considering sequence information. As observed in Tables 1 and 2, the MAPs of CM, BoW and CS are also improved by 20.3%, 16.1% and 16.3% respectively when sequence information is used. This demonstrates the advantage of using video sequence compared with static keyframes for event detection. The improvement is mainly due to two aspects. First, more visual information is employed when more frames are used. This improves the probabilities to make correct detections. Second, EMD matching can, if not perfectly, model some sequential information in the video clip, which proves to be useful in event detection. Although limited information is used, keyframe-based approaches can still achieve similar performance compared with sequence based approaches for those events that are not motion-intensive or do not show explicit motion patterns such as *People_Crying* and *Singing*. This is because the recognition of these events does not benefit much from the additional motion information used in sequence based approaches.

Among all features used in keyframe based approaches, as seen in Table 2, CM using the global visual feature of the keyframe gets the lowest AP. The local feature BoW again demonstrate its ability to discriminate different events as in concept detection [13] by achieving an improvement of 20.5% compared with CM. This is because the global features can be easily affected by noise in unconstrained environments and is thus weak in describing concepts/events. In contrast, local feature is capable of recognizing different concepts more effectively by extracting more discriminative local information. While using the expanded BoW (EBoW), the MAP is further improved by 7.2%.

Table 2: Average Precision (%) of event detection by keyframe based approaches using different features. CM: Color Moment; BoW: Bag-of-Visual-Words; EBow: Expanded BoW by visual relatedness; CS: Concept Score; MAP: Mean Average Precision.

Event	CM	BoW	EBoW	CS
Exiting_Car	16.9	20.1	**21.9**	18.6
Handshaking	6.7	8.8	9.4	**9.5**
Running	40.7	49.4	**52.3**	47.9
Demonstration_Protest	24.3	28.1	**29.9**	26.8
Walking	18.5	26.6	**27.8**	25.7
Riot	10.4	14.1	**14.7**	11.3
People_Marching	17.4	21.4	**23.3**	20.5
Dancing	12.3	12.8	13.9	**15.2**
Election_Campaign_Greeting	7.7	8.2	8.9	**9.2**
Shooting	9.3	8.7	9.4	**9.8**
Airplane_Flying	11.9	14.7	**15.6**	14.2
Street_Battle	18.6	21.8	**23.4**	20.7
People_Crying	4.8	7.1	7.9	**8.4**
Singing	9.2	9.8	11.2	**12.7**
MAP	14.91	17.97	**19.26**	17.89

7. CONCLUSION

In this paper we address event detection in unconstrained videos by proposing a new motion feature namely ERMH-BoW. While most existing features remain weak in event description, we show ERMH-BoW can completely and honestly describe an event. In ERMH-BoW, we adopt bag-of-visual-words to represent *what* aspect of an event, and capture the object activities or *how* aspect of the event by the relative motion histograms between visual words. The derived feature ERMH-BoW can thus provide a complete description of an event by closely integrating *what* and *how* aspects. Instead of motion distribution which is noisy, we employ local motion information of visual word, which is more discriminative by capturing the motion of interest for effective event detection. To address the motion distortion problem caused by various camera movement, motion relativity is employed. We demonstrate that relative motion is able to present the object activities in an event regardless of various camera movement. As we know, this is the first work to seriously study the motion distortion problem in event detection. In addition, we also propose to expand the motion histogram of visual words to alleviate the correlation problem by visual relatedness calculated from a visual word ontology. Our experiments show that ERMH-BoW can significantly improve the MAP of event detection compared

with existing features. Due to the use of sequence information, the extraction of ERMH-BoW is naturally slower than KF-based extraction. The speed is largely dependent on the number of frames being processed in the video examples. In our current implementation, the feature extraction and learning speed is approximately 10 times slower than that using only single keyframe. In the future, we will research efficient model for the utilization of motion relativity and visual relatedness for video event detection.

Acknowledgement

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118906) and a grant from City University of Hong Kong (Project No. 7002241).

8. REFERENCES

[1] E. Ardizzone, M. L. Cascia, A. Avanzato, and A. Bruna, "Video Indexing Using MPEG Compensation Vectors", *IEEE Int. Conf. on Multimedia Computing and Systems*, vol. 2, 1999.

[2] Oren Boiman and Michal Irani, "Detecting Irregularities in Images and in Video", *Int. Conf. on Computer Vision*, 2005.

[3] J. Cao *et. al*, "Intelligent Multimedia Group of Tsinghua University at TRECVID 2006", *TRECVID Workshop*, 2006.

[4] M. T. Chan, A. Hoogs, J. Schmiederer, and M. Petersen, "Detecting Rare Events in Video Using Semantic Primitives with HMM", *Int. Conf. on Pattern Recognition*, 2004.

[5] T. Chua, S. Neo, Y. Zheng, H. Goh, X. Zhang, S. Tang, Y. Zhang, J. Li, J. Gao, H. Luan, Q. He, and X. Zhang, "TRECVID 2007 Search Tasks by NUS-ICT", *TRECVID Workshop*, 2007.

[6] James W. Davis, "Hierarchical Motion History Images for Recognizing Human Motion", *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[7] L. Davis, S. Fejes, D. Harwood, Y. Yacoob, I. Hariatoglu, and M. J. Black, "Visual Surveillance of Human Activity", *Asian Conf. on Computer Vision*, 1998.

[8] Y. F. Day, S. Dagtas, M. Iino, A. Khokhar, and A. Ghafoor, "Object-Oriented Conceptual Modeling of Video Data", *Int. Conf. on Data Engineering*, 1995.

[9] S. Ebadollahi, L. Xie, Shih-Fu Chang, and J. R. Smith, "Visual Event Detection Using Multi-Dimensional Concept Dynamics", *IEEE Int. Conf. on Multimedia and Expo*, 2006.

[10] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance", *IEEE Int. Conf. on Computer Vision*, 2003.

[11] S. Gong and T. Xiang, "Recognition of Group Activities using Dynamic Probabilistic Networks", *IEEE Int. Conf. on Computer Vision*, 2003.

[12] A. Haubold and M. Naphade, "Classification of Video Events using 4-dimensional time-compressed Motion Features", *ACM Int. Conf. on Image and Video Retrieval*, 2007.

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

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

[15] Y. G. Jiang, C. W. Ngo, and J. Yang, "VIREO-374: LSCOM Semantic Concept Detectors Using Local Keypoint Features", http://vireo.cs.cityu.edu.hk/research/vireo374/.

[16] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", *Proc. of ROCLING X*, 1997.

[17] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection using Volumetric Features", *Int. Conf. on Computer Vision*, 2005.

[18] T. K. Kim, S. F. Wong, and R. Cipolla, "Tensor Canonical Correlation Analysis for Action Classification", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[19] Ivan Laptev and Tony Lindeberg, "Space-time Interest Points", *IEEE Int. Conf. on Computer Vision*, 2003.

[20] H. Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *Int. Conf. on Computer Vision*, 2007.

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

[22] B. D. Lucas and T. Kanade, "An Interative Image Registration Technique with an Application to Stero Vision", *Int. Joint Conf. on Artifical Intelligence,* pp. 121-130, 1981.

[23] Lihi Zelnik-Manor and Michal Irani, "Statistical Analysis of Dynamic Actions", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, 2006.

[24] Y. F. Ma and H. J. Zhang, "Motion Pattern-Based Video Classification and Retrieval", *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, pp. 199-208, 2003.

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

[26] C. W. Ngo, Y. Jiang, X. Wei, F. Wang, W. Zhao, H. Tan, and X. Wu, "Experimenting VIREO-374: Bag-of-Visual-Words and Visual-Based Ontology for Semantic Video Indexing and Search", *TRECVID Workshop*, 2007.

[27] J. C. Niebles, H. Wang, and Fei-Fei Li, "Unsupervised Learning of Human Action Categories Using spatial-Temporal Words", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[28] Y. Peng and C. W. Ngo, "EMD-based Video Clip Retrieval by Many-to-Many Matching", *Int. Conf. on Image and Video Retrieval*, 2005.

[29] E. Shechtman and M. Irani, "Space-Time Behavior Based Correlation", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[30] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, "The Function Space of an Activity", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[31] D. Xu and Shih-Fu Chang, "Visual Event Recognition in News Video using Kernel Methods with Multi-Level Temporal Alignment", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[32] Y. Yacoob and M. J. Black, "Parametrized modeling and recognition of activities", *Computer Vision and Image Understanding*, vol. 73, no. 2, 1999.

[33] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[34] D. Zhang, D. G. Perez, S. Bengio, and I. McCowan, "Semi-supervised Adapted HMMs for Unusual Event Detection", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[35] H. Zhong, J. Shi, and M. Visontai, "Detecting Unusual Activity in Video", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.

[36] DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, "Revision of LSCOM Event/Activity Annotations", Columbia University ADVENT Technical Report#221-2007-7, Dec 2006.

[37] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid.