Detecting Video Events Based on Action Recognition in Complex Scenes Using Spatio-Temporal Descriptor

Guangyu Zhu[1], Ming Yang[2], Kai Yu[2], Wei Xu[2], Yihong Gong[2]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[2] Department of Information Analysis, NEC Laboratories America, CA 95014, U.S.
gyzhu@jdl.ac.cn, {myang, kyu, xw, ygong}@sv.nec-labs.com

ABSTRACT
Event detection plays an essential role in video content analysis and remains a challenging open problem. In particular, the study on detecting human-related video events in complex scenes with both a crowd of people and dynamic motion is still limited. In this paper, we investigate detecting video events that involve elementary human actions, *e.g. making cellphone call*, *putting an object down*, and *pointing to something*, in complex scenes using a novel spatio-temporal descriptor based approach. A new spatio-temporal descriptor, which temporally integrates the statistics of a set of response maps of low-level features, *e.g.* image gradients and optical flows, in a space-time cube, is proposed to capture the characteristics of actions in terms of their appearance and motion patterns. Based on this kind of descriptors, the bag-of-words method is utilized to describe a human figure as a concise feature vector. Then, these features are employed to train SVM classifiers at multiple spatial pyramid levels to distinguish different actions. Finally, a Gaussian kernel based temporal filtering is conducted to segment the sequences of events from a video stream taking account of the temporal consistency of actions. The proposed approach is capable of tolerating spatial layout variations and local deformations of human actions due to diverse view angles and rough human figure alignment in complex scenes. Extensive experiments on the 50-hour video dataset of TRECVid 2008 event detection task demonstrate that our approach outperforms the well-known SIFT descriptor based methods and effectively detects video events in challenging real-world conditions.

Categories and Subject Descriptors
I.4.8 [**Scene Analysis**]: *Motion*, *Object Recognition*, *Tracking*. I.2.10 [**Vision and Scene Understanding**]: *Video Analysis*.

General Terms
Algorithms, Performance, Experimentation.

Keywords
Motion Representation, Action Recognition, Event Detection, Semantic Analysis.

1. INTRODUCTION
With the explosive growth in the amount of digital videos and rapid advance in the computing power of computers, the management and retrieval of video data has been actively studied in the past few years. Event detection is particularly crucial for understanding semantic concepts of interest in videos for intelligent management and advanced retrieval purposes. Therefore, extensive research efforts have been devoted to event-based video analysis [1]-[7][13]-[20].

Considering elementary semantic concepts involved in video events, a distinction can be generally made between object-oriented (or static-concept) events and action-oriented (or dynamic-concept) events. The events comprising the concepts like *Cityscape* and *Boatship* are object-oriented in the sense that they are primarily concerned with the presence of particular objects in a video stream. In the high-level feature extraction task of annual TREC video retrieval evaluation (TRECVid) [9], the benchmark of annotated video corpus is provided to researchers for detecting a large set of object-oriented events. In contrast, the action-oriented events, such as *People-calling-cellphone* (CellToEar), *People-dropping-something* (ObjectPut) and *People-pointing-something* (Pointing), involve the semantic concepts that are exclusively related with specific actions performed in a video stream.

Action-oriented event detection is an important component for many intelligent video management applications especially in surveillance video analysis for security [4], sports video analysis for labeling and searching [5], and online video repository searching and mining [10]. Consequently, there exists a compelling demand for investigating effective and efficient approaches for action-oriented event detection in videos. Moreover, a large amount of human-related action-oriented events occur in complex scenes where the same type of actions may exhibit enormous variations due to clutter background, different viewpoints and many other factors (*e.g.* human-body occlusions and low-resolution videos) in unconstrained real-world environment. To the best of our knowledge, the related work on detection of action-oriented video events in real-world conditions is still limited [3][5]-[7].

This line of research suffers from a lack of standard benchmark video dataset which supplies sufficient clearly defined video events together with ground truth annotations in unconstrained real-world environment. Most of the existing datasets for action recognition or event detection, *e.g.*, the KTH dataset [8], were recorded in a controlled setting with slight camera motion and clean background. Fortunately, TREC video retrieval evaluation 2008 [9] launched a new task of action-oriented event detection, which made the largest effort to bridge the research efforts and the challenges in real-world conditions by providing an extensive 99-hour airport surveillance video dataset. This task is intended to help promote the technology development for event detection, especially leveraging machine learning techniques to detect a pre-

MM'09, October 19–24, 2009, Beijing, China.

defined set of events. Figure 1 illustrates some samples of the required events in this dataset. The highly crowded scenes, the extremely cluttered background, and the versatility of execution styles of the same actions make action-oriented event detection a formidably challenging task. Such a standard benchmark dataset is indispensable for the development of research. Since the task of event detection in TRECVid evaluation was launched, it has attracted significant attention from academia as well as industry.

(a) CellToEar (b) CellToEar

(c) ObjectPut (d) Pointing

Figure 1. Samples of some pre-defined events in different cameras of the TRECVid 2008 event detection dataset.

In this paper, we investigate action-oriented event detection in challenging complex scenes where the actions of interest are masked by the activities in a dynamic and crowded real-world environment. Generally, the action-oriented events of interest are application dependent. We consider the elementary actions that are some articulated motion of a single human body which cannot be easily decomposed to simpler actions. In particular, we focus on three events *CellToEar*, *ObjectPut* and *Pointing* which are defined in the TRECVid 2008 event detection task. The detailed descriptions of these three events are listed in Table 1. In our work, as the major component of event detection, an action recognition method is developed utilizing bag-of-words models of novel spatio-temporal descriptors to train support vector machine (SVM) classifiers at multiple spatial pyramid levels. Further, resorting to a temporal filtering strategy, the event sequences are segmented from the video stream.

Table 1. Description of three events of interest

Event	Description
CellToEar	Someone puts a cell phone to his/her ear
ObjectPut	Someone drops or puts down an object
Pointing	Someone points something

The novelty and contributions of this paper are summarized as follows. 1) We propose a novel spatio-temporal descriptor, the temporally integrated spatial response descriptor (henceforth abbreviated as TISR), which integrates the temporal statistics of a set of response maps of low-level image features, *e.g.* image gradient and optical flow fields, in a space-time cube. Compared with the existing local descriptors like SIFT [31], this kind of descriptors can delineate the local patterns of image patches in terms of their appearance and motion characteristics and are robust to variations and deformations of objects in real-world conditions. 2) The bag-of-words (BoW) method combined with the spatial pyramid technique is employed to generate the compact feature representations of human figures in action-oriented events. These representations are insensitive to rough human figure alignment as well as some influence factors such as partial occlusions, background clutter and pose changes. 3) A Gaussian kernel based temporal filtering method is utilized to segment the event sequences from a video stream by considering their temporal consistency. 4) As demonstrated by extensive experiments on the TRECVid 2008 event detection dataset, our approach using the TISR descriptors no matter extracted from image gradient or optical flow fields or both of them consistently outperforms the method using the well-known SIFT descriptors in most of the cases for the task of action-oriented video event detection.

The rest of the paper is organized as follows. Section 2 reviews the related work of video event detection and human action recognition. In Section 3, the overview of the proposed event detection approach is presented. In Section 4, we introduce the novel action recognition approach based on the new TISR descriptor. Section 5 describes the temporal filtering strategy for event segment detection from a video stream. Experimental results are reported and analyzed in Section 6. Finally, we conclude the paper with future work in Section 7.

2. RELATED WORK

As a sub-area of event-based video analysis, human-related action-oriented event detection shares the common procedures of video event detection including extracting relevant features and making detection decision, yet it mainly leverages action recognition techniques as the cornerstone for event detection. In this section, we present a brief review of the state-of-the-art regarding the research of video event detection and human action recognition in real-world conditions.

2.1 Video Event Detection

Event detection for various applications has been studied in [11]-[20]. Detailed surveys on this topic can be found in [11] and [12]. The conventional procedure in the existing event detection methods can be generally divided into two steps [13]: 1) generating video content representation exploiting various properties extracted from raw video stream and 2) making detection decision using certain classification techniques.

For video content representation, the existing studies employ the properties extracted from video stream including visual features [3], audio features [14], text features [15][16] and the combination of the multimodal features [2][13] to facilitate accurate detection. On the other side, detection decision making plays a very important role in determining the final performance. Various classification techniques have been applied to discover the event patterns from large scale video sets. One of the examples is the work proposed by Xie *et al*. [17] in which the hidden Markov model framework was developed to discover the patterns in soccer video. Xu *et al.* [3] developed a discriminative kernel-based visual event detection method via special multilevel alignment. In [18], Shyu *et al*. recently proposed a subspace based data mining framework

for event detection which includes three components: video pre-processing, distance based data mining and rule based data mining. In addition, C4.5 decision tree [19] and SVM [20] have been widely used as the classifiers for the decision of event detection.

2.2 Human Action Recognition in Real-World Conditions

Action recognition is one of the most challenging problems in the area of video analysis. In [21], Turaga *et al*. presented a recent survey of the major approaches pursued over the last two decades. To make this problem more tractable, most of existing approaches made simplified assumptions, *e.g.* clean background or little viewpoint changes, and were designed for constrained conditions in laboratories or studios. For real-world deployment, action recognition systems however need to be robust against numerous factors, *e.g.* noise, occlusions or shadows, in unconstrained real-world conditions.

A few work [3][5]-[7] attempted to perform human action recognition in real-world conditions for applications of video event detection. In [3], Xu *et al*. systematically studied the problem of visual event recognition in unconstrained news videos. The earth mover's distance within the bag-of-words method is utilized to evaluate the similarity among video clips. By fusing the information from a multi-level similarity pyramid, the recognition is conducted in the framework of temporally aligned pyramid matching. In [5], Zhu *et al*. proposed a new action descriptor based on the insight of treating optical flow field as spatial patterns of noisy measurements instead of precise pixel displacements at points. Using those action descriptors, the actions of players in the far-view shots within broadcast tennis videos are recognized. Ke *et al*. [6] employed a combination of shape and flow features to recognize the actions in the events of interest in cluttered scenes. Laptev *et al*. proposed a multimodality based action classification approach [7]. The movie scripts are first employed to address the problem of automatic human action annotation. By extending the idea of local space-time features, space-time pyramids and multichannel non-linear SVM, Laptev's method for action classification achieved good performance on a movie dataset.

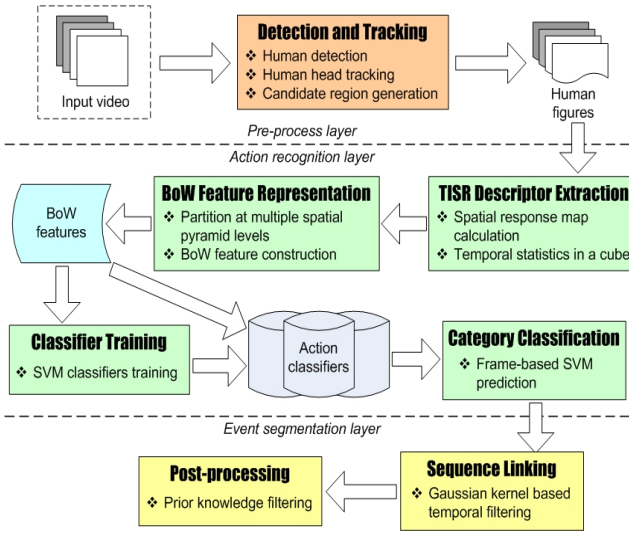


Figure 2. Flowchart of the proposed action-oriented event detection method based on human action recognition.

3. OUR APPROACH

Human action-oriented events essentially involve sequences of specific human postures evolving in video streams, which typically demonstrate considerable variations in both spatial and temporal domains. For instance, the action of dropping a bag may appear quite different in monocular videos from various view angles and the durations may also vary case by case. Thus, robust action representations that are invariant or at least tolerant to both spatial and temporal variations are indispensable for detecting action-oriented video events. Our approach strives to capture the characteristics of individual actions by extracting dense spatio-temporal descriptors and representing actions by bag-of-words features of these salient descriptors. The proposed TISR descriptor fuses the temporal statistics of a few response maps of different low-level image features in a space-time cube. With a visual vocabulary of the TISR descriptors, the BoW histogram features are able to tolerate spatial and temporal variations. Then, we segment event sequences spatially by taking advantage of human detection and tracking and temporally by Gaussian kernel filtering.

In general, the procedure of an action-oriented video event detection approach can be divided into three layers: 1) a pre-process layer in which the candidate regions of interest are located; 2) an action recognition layer in which the action category is recognized by classifying a compact representation extracted from the candidate region; 3) an event segmentation layer in which the action recognition results are linked by temporal filtering and cleaned if prior knowledge about the scene is available. Following this paradigm, Figure 2 illustrates the flowchart of our approach. Given the input video sequence, we first locate the candidate regions to analyze by human head detection and tracking algorithms. To obtain the human figure, an enlarged region around the tracked head is cropped as the input to the action recognition module. Then, for each human figure, the dense TISR descriptors are extracted from the response maps of image gradients and optical flows in a volumetric cube. We construct bag-of-words (BoW) features by measuring the frequencies of quantized descriptors with a visual vocabulary at multiple spatial pyramid resolution levels. Afterwards, the action category is classified by fusing the classification results of SVM classifiers at all spatial pyramid levels. Based on the frame-based recognition results, a temporal filtering using Gaussian kernel is employed to segment the event sequences from video stream. In post-processing, scene prior knowledge is used to reduce some false alarms.

4. ACTION RECOGNITION BASED ON SPATIO-TEMPORAL DESCRIPTOR

An effective human action recognition method is the core of the action-oriented event detection. Three key issues need to be addressed: 1) what spatio-temporal features relevant to actions shall be extracted, 2) how to organize these features to represent human figures, and 3) how to classify the human figures to different action categories. In this paper, we extract the dense TISR descriptors to build BoW features and utilize SVM classifiers at multiple spatial pyramid resolution levels to address these problems.

4.1 Spatio-Temporal Descriptor Extraction

The proposed spatio-temporal descriptor which is named as temporally integrated spatial response (TISR) is extracted from pixel-level vector fields of low-level image features, *e.g.* image gradient and optical flow, within a space-time cube. First, a set of response

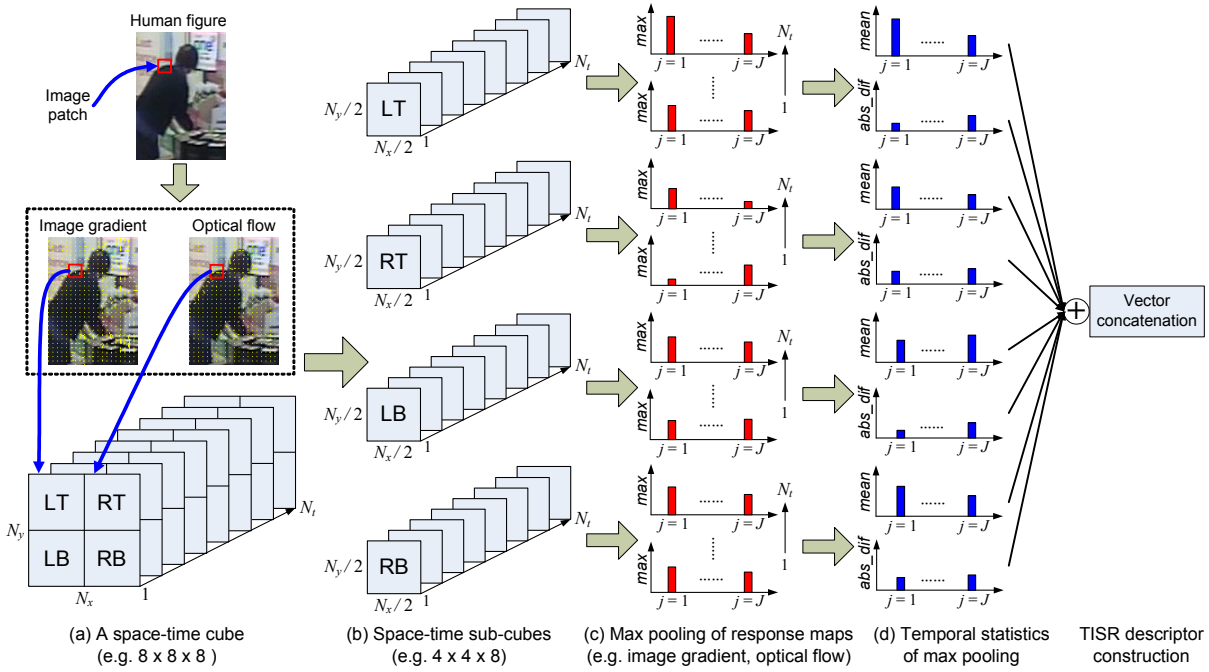


Figure 3. Algorithm of the TISR descriptor extraction. The descriptor is extracted from the space-time cube corresponding to the image patch drawn as the red rectangle based on low-level image gradient and optical flow fields.

maps of different preferred directions and magnitudes are calculated based on the vector fields. Then, the TISR descriptor is constructed by integrating the temporal statistics on the spatial maximum pooling results of individual response maps in more elaborately partitioned spatial-temporal sub-cubes. The entire procedure of TISR descriptor extraction is illustrated in Figure 3. In the process of extraction, the spatial partition and maximum pooling extract the salient appearance information while the integration of some statistics of maximum pooling results along the time axis reveals the properties of motion patterns. Therefore, the TISR descriptor can well delineate the local appearance and motion characteristics in a space-time cube.

4.1.1 Response Map Calculation from Low-level Appearance and Motion Features

The TISR descriptor can be extracted from arbitrary vector fields of low-level image features. In our implementation, we use image gradient and optical flow for efficient processing. Given the input frame I_t at time t, the 2D vector fields of image gradient G_t and optical flow F_t are computed using Sobel operator [34] and Horn-Schunck algorithm [35] which shown empirical good performance.

Motivated by the work in [32][5], half-wave rectification and Gaussian smoothing are applied to mitigate the noise in the vector fields of image gradient and optical flow. Half-wave rectification can make the data sparse and avoid the cancellation of vectors with opposite directions during smoothing, so that significant Gaussian smoothing can be applied to reduce the amount of noise. The process is shown in Figure 4. Let **VF** represents a vector field of either image gradient or optical flow, the magnitudes of **VF** are first thresholded to reduce the influence of too small and too large edges or motion probably due to noise. Then, the **VF** is split into 2 scalar fields corresponding to the horizontal and vertical components VF_X and VF_Y, which are then half-wave rectified into 4 non-negative channels VF_X^+, VF_X^-, VF_Y^+, and VF_Y^-, where they satisfy $VF_X = VF_X^+ - VF_X^-$ and $VF_Y = VF_Y^+ - VF_Y^-$. Each of these 4 fields is smoothed by a Gaussian filter. Thus the noise in the original field is largely reduced and the refined vector fields are obtained.

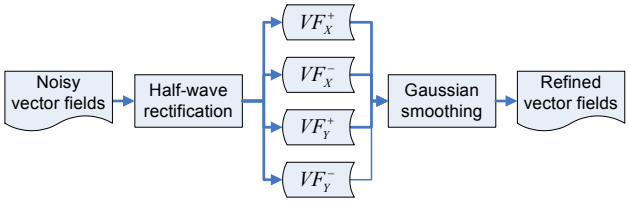


Figure 4. Half-wave rectification and Gaussian smoothing for noise reduction of image gradient and optical flow fields.

Thereafter, for the vector fields G_t and F_t, a set of response maps of different directions and magnitudes for each pixel in the fields are computed by applying the following equation

$$R(\theta, V \mid \theta_p, V_p) = R_\theta(\theta, \theta_p) \cdot R_V(V, V_p), \qquad (1)$$

where θ and V denote the direction and the magnitude of the vector at each pixel position in the field, θ_p and V_p indicate the preferred direction and magnitude of each response map. We employ 8 directions $\theta_p = \{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$ and 2 magnitudes in the response map calculation. The two preferred magnitudes are empirically set to be $V_p = \{75, 150\}$ and $V_p = \{2.5, 5\}$ for image gradient and optical flow, respectively. For image gradient, the function $R_\theta(\bullet)$ and $R_V(\bullet)$ for response calculation are defined in Eq. (2) using triangular functions [25].

$$R_\theta^{IG}(\bullet)=R_V^{IG}(\bullet)=Tr(x,x_p)=\begin{cases}1-\left|\dfrac{x-x_p}{\omega}\right|, & \text{if } \left|\dfrac{x-x_p}{\omega}\right|<1\\ 0, & \text{otherwise}\end{cases}, \quad (2)$$

where x and x_p represent the inputs θ (or V) and θ_p (or V_p), ω is a scale parameter of the function which is set to be 45° for direction and 75 for magnitude. On the other hand, the response functions of optical flow are defined as

$$\begin{aligned}R_\theta^{OF}(\theta,\theta_p)&=\left\{0.5\times\left[1+\cos\left(\theta-\theta_p\right)\right]\right\}^q\\ R_V^{OF}(V,V_p)&=\exp\left(-\left|V-V_p\right|\right)\end{aligned}, \quad (3)$$

which are inspired from biological study in [26] and the parameter $q=2$ controls the width of the tuning curve [26]. The empirical parameters are selected according to the overall statistics of image gradients and optical flows. Taking all the possible combination of θ_p and V_p, we obtain a set of 16 types of response maps for G_t and F_t, respectively.

The extraction of the response maps over the discrete directions and magnitudes can further reduce the influence of noise in the low-level features. More important, different from quantizing the response of a pixel to a single preferred direction and magnitude, the pixel response in our method contributes to multiple response maps, thus the calculation of the set of response maps is kind of soft quantization and preserves the relative differences among different directions and magnitudes.

4.1.2 Descriptor Extraction from Space-Time Cubes

After calculating the set of response maps of each frame, we crop a space-time cube based on the candidate region to analyze (*e.g.* a human figure for action-oriented event detection) and extract dense TISR descriptors. As shown in Figure 5, given an image region of interest (the red rectangle region), its space-time cube is constructed by concatenating the image regions (the blue rectangle regions) at the same coordinates in the successive video frames along the time axis. In the implementation, we extract such space-time cubes for each type of response map obtained in Section 4.1.1.

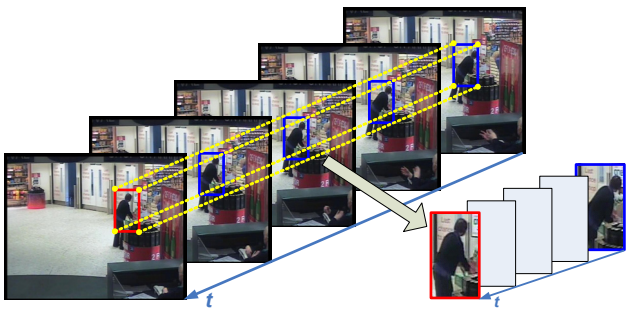


Figure 5. Construction of the space-time cube structure of a human figure from a video stream.

As shown in Figure 3(a), given a human figure HI to analyze in frame t, the image patches with the spatial dimension $N_x \times N_y$ are densely sampled within HI. The corresponding space-time cubes are constructed by concatenating image patches from N_t frames in the video stream. In our work, we set $N_x = N_y = 8$ or 16 and $N_t = 8$. To preserve the spatial layout information, each cube is divided into several non-overlapping sub-cubes. Here, 2×2 grid style is employed, in which the sub-cubes are annotated as *LT* (left-top), *RT* (right-top), *LB* (left-bottom) and *RB* (right-bottom) as shown in Figure 3(b). The sub-cube is used to calculate the temporal statistics of each response map and then these statistics in the 4 sub-cubes are concatenated to form the TISR descriptor of an $N_x \times N_y \times N_t$ space-time cube. Since the same procedure is applied to sub-cube, we ignore the index of sub-cube to make the description more concise in the following description.

The sub-cubes of different response maps are denoted by $RI_{i,j}^{IG}$ and $RI_{i,j}^{OF}$ regarding to low-level image gradient and optical flow calculated using Eq. (2) and Eq. (3) respectively, where $i=1,\ldots,N_t$ represents the temporal index and $j=1,\ldots,J$ represents the index of the combination set for all the preferred orientations θ_p and magnitudes V_p. For example, $j=1$ indicates $\theta_p=0°$ with $V_p=75$ and $j=16$ indicates $\theta_p=315°$ with $V_p=150$ for image gradient. For each response map $RI_{i,j}^{*}$ where the symbol star represents image gradient (*IG*) or optical flow (*OF*), the maximum pooling operation is conducted to obtain the spatial local maximum $h_{i,j}^{*}$. Figure 3(c) shows the max-pooling results of J response maps at each time instance, which are denoted by H_i^{*}, $i=1,\ldots,N_t$,

$$H_i^{*}=\left\{h_{i,j}^{*}\mid h_{i,j}^{*}=\max\{RI_{i,j}^{*}\},j=1,\ldots,J\right\}. \quad (4)$$

The existing work [27] has demonstrated that max-pooling is a good means to increase the tolerance to the variance of local transformations and object deformations and enhance the robustness to background clutter.

Based on the max-pooling results, the mean m_j^{*} and the average of absolute difference a_j^{*} over N_t frames are calculated for J types of response maps as shown in Figure 3(d).

$$m_j^{*}=\frac{\sum_{i=1}^{N_t}h_{i,j}^{*}}{N_t}, \quad (5)$$

$$a_j^{*}=\frac{\sum_{i=2}^{N_t}\left|h_{i,j}^{*}-h_{i-1,j}^{*}\right|}{N_t-1}. \quad (6)$$

Such two statistics integrates the spatial max-pooling results in the temporal domain, which is capable of summarizing the temporal characteristics or motion patterns of local image appearances. The descriptor is obtained by concatenating all the statistics m_j^{*} and a_j^{*} of J types of response maps of both image gradient and optical flow as following.

$$\left[m_1^{*},a_1^{*},m_2^{*},a_2^{*},\ldots,m_J^{*},a_J^{*}\right]. \quad (7)$$

Finally, the TISR descriptor of the entire cube is the concatenation of the descriptors extracted from all the sub-cubes. For the configuration in our work, the descriptor is a 256-dimension vector, *i.e.* $16\times2\times4\times2$, where 16 indicates the number of different types of response maps, 2 statistics, 4 sub-cubes and 2 types of low-level features, *i.e.* image gradient and optical flow.

Consequently, given a human figure HI, we can compute a set of spatio-temporal descriptors $D=\{d_1,\ldots,d_M\}$, where M is the

number of descriptors extracted to represent the characteristics of the figure in terms of appearance and motion patterns. This set of descriptors will be used to generate a compact feature representation using the bag-of-words method.

4.2 Bag-of-Words Feature Representation

Based on the TISR descriptors, the bag-of-words histogram feature is generated to describe a human figure as a compact representation. The bag-of-words method has recently attracted increasing research attention since its success in object categorization. The procedure of BoW feature generation has two steps: visual vocabulary construction and feature vector representation.

A visual vocabulary is constructed by clustering the TISR descriptors and treating each cluster as a single visual word. An issue of vocabulary construction is how to determine the adequate number of visual words in the vocabulary. A small vocabulary may lack sufficient discriminative power while a large vocabulary, on the other hand, may be less generalized. There is no theoretical guide for the determination of vocabulary size. According to our experiments on the TRECVid 2008 dataset, we employ a vocabulary with 512 visual words for the TISR descriptors.

In the process of building the BoW feature for a human figure, a histogram is obtained by quantizing each TISR descriptor to a visual word and counting the frequency of each visual word in visual vocabulary. The basic idea of bag-of-words method is to depict the image as an orderless collection of local descriptors with the result that it completely disregards the spatial locations and layout of the descriptors in the image. Recently, the spatial pyramid matching (SPM) technique [28] follows the strategy of "subdivide and disorder" to compensate this loss. In our work, given a spatial pyramid resolution level with a subdivision style, a histogram feature is calculated in each local partition region using the bag-of-words routine. The final feature representation of a candidate region or a human figure is given by the concatenated vector of multiple BoW histograms extracted from all the local partitions. Figure 6 illustrates the subdivision styles of spatial pyramid levels used in our work. From resolution 1 to L, the grid at level l has 2^{l-1} partitions along each dimension.

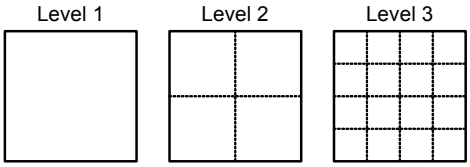


Figure 6. The spatial pyramid partition styles for different levels, which are 1x1, 2x2 and 4x4 respectively.

Given the set of TISR descriptors of a human figure $D = \{d_1,\ldots,d_M\}$, the BoW feature is calculated on each level and represented as $F = \{f_1,\ldots,f_L\}$, where $L = 3$ is the resolution levels in this paper. Using the vocabulary with 512 visual words, the feature dimensions for each level are 512, 2048 and 8192 for $F = \{f_1,\ldots,f_L\}$, respectively.

4.3 Action Classification Using SVM

We formulate action recognition as a multiclass classification task. Various supervised learning algorithms can be employed to train the action classifiers. We employ the widely used SVM classifiers [29] in our approach due to its superb generalization capability to unseen test data as well as less parameters to tune. In addition, the existing work [30] has demonstrated that bag-of-words features achieved good classification performance using SVM classifiers.

The input of the action classification module is the BoW feature $F = \{f_1,\ldots,f_L\}$ extracted from L spatial pyramid levels for a human figure. For each action category, L SVM classifiers $SC = \{C_1,\ldots,C_L\}$ are trained in which the classifier C_l employs the BoW feature f_l at resolution level l as the input. One-against-all classification scheme is employed. The linear kernel is utilized to map the training vectors into a high dimensional feature space for classification. Compared with other types of kernel functions, the linear kernel has the advantage of lower computational complexity which is more suitable for the huge dataset of TRECVid event detection task.

The output of the linear SVM is the distance that the input feature is away from the classification boundary defined by support vectors. To convert the distance to a likelihood value, the sigmoid function [33] is employed to transform the classification output to the likelihood that the human figure is performing a certain action. The transformation is defined as

$$T(e_l) = \frac{1}{1+\exp(\alpha \cdot e_l + \beta)}, \tag{8}$$

where $e_l = C_l(f_l)$ is the distance outputted by the C_l classifier with the input feature f_l, $T(e_l)$ is the corresponding probabilistic likelihood, the parameters α and β are empirically set as 1.0 and 0.0 in our experiments, respectively.

Motivated by the weighting strategy of spatial pyramid matching, for the human figure with the feature $F = \{f_1,\ldots,f_L\}$, the final classification likelihood E_r corresponding to the category r is given by the weighted sum of all the likelihoods obtained at different spatial pyramid resolution levels.

$$E_r = \frac{1}{Z}\sum_{l=1}^{L}\frac{1}{2^{L-l}}T(e_l), \tag{9}$$

where $Z = \sum_{l=1}^{L} 1/2^{L-l}$ is the normalization constant.

5. TEMPORAL FILTERING FOR EVENT SEQUENCE SEGMENTATION

Our classification results of human figures are primarily frame-based. To further segment the event sequences from the video stream, we employ a temporal Gaussian filtering to link and smooth the frame-based classification results. Afterwards, scene prior knowledge can be used to further remove some false alarms to enhance the final detection accuracy.

Gaussian filtering is essentially a low-pass filter using a Gaussian kernel. The purpose of the temporal Gaussian filtering is to suppress the noise in the frame-based classification results and link them to video event segments by taking account of the temporal consistency of the actions. Denote the Gaussian kernel function as $g(\bullet) \sim G(0,\sigma)$ where the deviation σ indicates the expected span of an action. Figure 7 illustrates the process of Gaussian

filtering. Given the event category r and the sequence of frame-based action recognition results $Q(t) = \{E_r(t_1), \ldots, E_r(t_n)\}$ which are on one trajectory, the Gaussian filtering is conducted by convoluting $Q(t)$ with the Gaussian kernel function

$$\bar{Q}(t) = \int_1^n Q(u) \cdot g(t-u)du \,. \tag{10}$$

As shown in Figure 7, after Gaussian filtering, the successive frames with the likelihoods higher than a pre-defined threshold (0.9 in the experiments) are extracted as the video event segments.

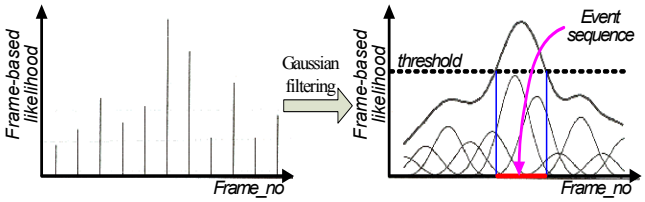
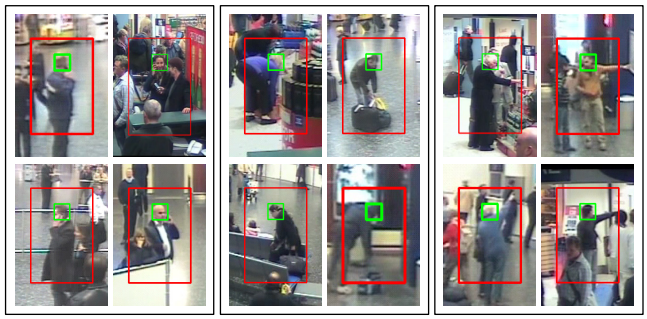


Figure 7. Event sequence segmentation using temporal Gaussian filtering.

After obtaining the event segments, scene prior knowledge, such as the ground plane homography or 3-D layout of the scene can be further leveraged to remove some false detections. In our implementation, we average all the human detections in the entire dataset and estimate a mask for active foreground regions for individual scenes. Some false detections occur outside the active foreground regions can be removed at the post-processing stage.

6. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed approaches, we performed thorough experiments on the TRECVid 2008 event detection dataset to detect 3 action-oriented events: *CellToEar*, *ObjectPut* and *Pointing*. The TRECVid event detection dataset was obtained from the Gatwick Airport which consists of 50-hour (5 days × 2 hours/day × 5 cameras) videos in the development set and 49-hour videos in the evaluation set. Our experiments were conducted on the development set since the ground truth annotations are available. For each video in the development set, there are about 190K frames with image resolution 720 × 576. The preliminary annotations of the occurrences of actions in the development set were provided by NIST. We further labeled the precise locations of persons performing the actions every 3 frames for training. Some positive training samples of the 3 events of interest are shown in Figure 8, where we observe large intra-class variations due to different viewpoints and the diverse ways people performing the same actions.



(a) Samples of *CellToEar* (b) Samples of *ObjectPut* (c) Samples of *Pointing*

Figure 8. Positive training samples in the TRECVid dataset.

To quantitatively evaluate the performance, we calculate the detection rate (DR) and false alarm (FA) for each event category, which are defined as follows.

$$DR = TP/(TP + FN)\,, \tag{11}$$

$$FA = FP/(FP + TN)\,, \tag{12}$$

where for each type of event, TP is the number of true positive instances, FN is the number of false negative instances, FP is the number of false positive instances, and TN is the number of true negative instances.

In the following sub-sections, we first present the results of human detection and tracking. Then, the frame-based performance of our method using the TISR descriptors extracted from different low-level features are compared with that of a frame based and a spatio-temporal method [36] using the SIFT descriptor [31]. Finally, the sequence-based performance is evaluated using the criteria of TRECVid 2008 event detection task.

6.1 Human Detection and Tracking

In the pre-processing layer, our action-oriented event detection method starts by human detection and tracking to locate candidate regions of interest. In our system, we employ a dedicated human detector based on convolutional neural networks (CNN) [22] and a tracker integrating multiple cues [23][24] to locate human heads. We bias to high detection rate rather than high precision. The overall detection rate of human including both detection and tracking results is tuned to about 80% with the precision 50%-60% approximately. Sample frames of detection and tracking results are shown in Figure 9. To obtain the human figure, the bounding box around the head region is enlarged to roughly contain the human body as shown in Figure 8.

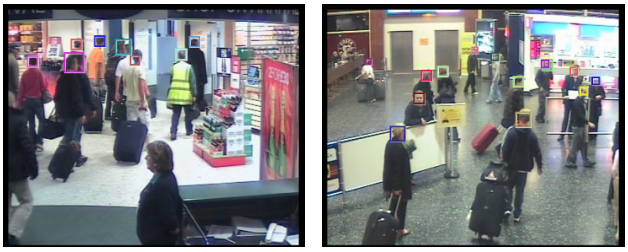


Figure 9. Samples of human detection and head tracking results where head locations are drawn as color rectangles.

6.2 Evaluation of Frame-based Performance

We first evaluate the TISR descriptor extracted from different low-level features in terms of the frame-based performance. The low-level features are image gradient, optical flow and the combination of both. The performance is compared with the well-known SIFT descriptor [31] which is acknowledged as one of the most powerful local feature descriptors and has achieved overwhelming successes in object categorization and recognition.. These evaluations demonstrate that 1) the TISR descriptor outperforms SIFT descriptor for the task of action-oriented event detection no matter using image gradients or optical flow, 2) our method can effectively combine multiple low-level features to improve the recognition performance and 3) our method outperforms the SIFT based spatio-temporal approach [36] on the TRECVid event detection dataset regarding the capacity for integration of spatial and temporal information.

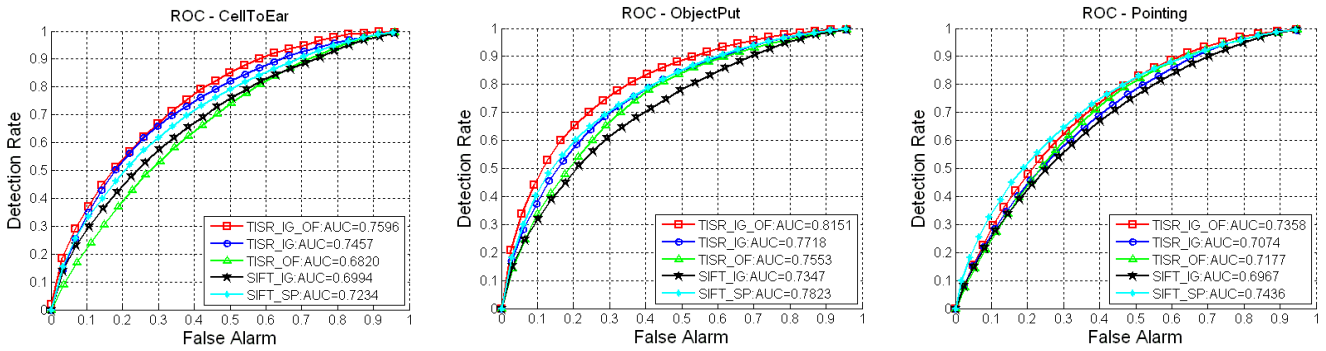


Figure 11. Frame-based ROC curves and AUC scores for the detection of CellToEar, ObjectPut and Pointing.

In the experiments, the positive training samples of an action are the frame-based labeled instances and the negative samples are the human detection and tracking outputs including both true and false detections. Table 2 shows the detail of the train samples, from which we can observe severe unbalance between the amounts of positive and negative samples. To mitigate the data unbalance, we perturb the positive samples to generate more instances for training. One positive sample is perturbed to 7 samples including 2 by zoom-in and zoom-out and 4 by shifting the figure center plus the original one. This also helps improve the generalization ability of the classifier.

Table 2. Detail of the training samples for action recognition

Event	# Positive sample	# Negative sample
CellToEar	2469	148640
ObjectPut	2974	
Pointing	10170	
Total	15613	148640

Since the videos were recorded on 5 different days, we therefore perform 5-fold cross-validation accordingly, which guarantees no identical samples appear in both training and testing sets. We adopt the one-against-all strategy to train separate classifiers for each action category. Then, we evaluate the frame-based classification results quantitatively and draw the average ROC curve with the average area under curve (AUC) score over 5 folds for each action. Greater the AUC score is, better the performance of the approach is.

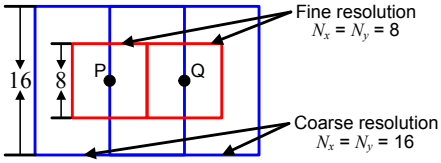


Figure 10. Multiscale resolution in the descriptor extraction for two image patches centered at *P* and *Q*.

To make the descriptors robust to the scale changes of the human action, we employ two spatial scales for extracting the dense TISR descriptors. As shown in Figure 10, the scale of a fine resolution is $N_x = N_y = 8$ and a coarse resolution is $N_x = N_y = 16$. The two adjacent image patches are non-overlapping at the fine resolution, while there is 50% percentage overlapping area at the coarse resolution. The descriptors of 75000 random sampled image patches including both scales from the whole dataset are used to construct a 512-word visual vocabulary by the K-means clustering algorithm. Then, this visual vocabulary is employed to generate the bag-of-words features at multiple spatial pyramid levels for each training sample as explained in Section 4.2.

For the comparison with the SIFT descriptor, the same training and testing datasets with the same evaluation settings are used for the SIFT based method. Basically, we substitute our descriptors by the SIFT descriptors in the proposed action recognition framework. Dense SIFT descriptors are extracted within the same cubes using 2 spatial scales and then the BoW features for $L = 3$ spatial pyramid resolution levels are built using a 512-word vocabulary to train the SVM classifiers. This approach is denoted as SIFT_IG. SIFT descriptors are extracted only from image gradients. For a fair comparison, we implement a simplified version of TISR descriptor that also uses the response maps of image gradient only. Thus, the dimensionality of this TISR descriptor is 128 the same as the SIFT. The approach using the simplified TISR descriptor is denoted by TISR_IG. Moreover, the methods using the TISR descriptors which exploit optical flow only and both of image gradient and optical flow are referred as TISR_OF and TISR_IG_OF, respectively. Furthermore, we compare with the SIFT based spatio-temporal approach [36] applied to the TREC-Vid 2008 event detection task. In this approach, the SPM features are constructed based on dense SIFT descriptors. Then, the statistics along the time axis are calculated using Eq. (5) and Eq. (6) from space-time cubes as shown in Figure 5 and fed to the SVM classifier. Because of the space limitation, please refer to [36] for the technical detail. This approach is denoted by SIFT_SP in our experiments. The ROC curves and the corresponding average AUC scores of different methods are illustrated in Figure 11.

The comparison of TISR_IG and SIFT_IG aims to evaluate the effectiveness of different descriptors to delineate spatio-temporal patterns within the BoW framework for action recognition. As shown in Figure 11, we can see that the TISR_IG achieves AUC scores 0.7457, 0.7718, and 0.7074 for the 3 action-oriented events *CellToEar*, *ObjectPut* and *Pointing*, respectively, which outperforms the SIFT_IG method by 0.031 in terms of the average AUC score. This verifies that our way to integrate the temporal statistics of spatial response maps of image gradients is more effective than the SIFT descriptor for the task of action-oriented event detection. In addition, the ROC curve and AUC score of the TISR_OF are also better than those of the SIFT descriptor based method. In the TISR_IG_OF method, the descriptor concatenates the temporal statistics of both image gradient and optical flow

thus has 256 dimensions. As shown in Figure 11, the AUC scores increase by 0.060 on average for 3 events compared with the SIFT descriptor based method. Also, the performance is better than either TISR_IG or TISR_OF. This result demonstrates that integrating more local features within the proposed TISR descriptor can effectively improve the event detection performance. The combination of image gradient and optical flow delineates the action patterns more comprehensively from both appearance and motion perspectives. The TISR_IG_OF method outperforms the SIFT_SP method for *CellToEar* and *ObjectPut* by 0.0362 and 0.0328 in AUC scores, respectively. The performance of *Pointing* is comparable to that of SIFT_SP. Such results demonstrate that the TISR descriptor is more effective in extracting spatio-temporal characteristics for action representation in real-world conditions.

(a) an FD for CellToEar (b) an FD for ObjectPut (c) an FD for Pointing

Figure 12. Samples of false detections (FD).

We still observe quite a few false alarms in the detection results. Some typical false detections are shown in Figure 12. The reasons for the incorrect detection are on two-fold. 1) The semantic gap between motion patterns and actions: some false detections are reasonable in the sense that the subtleties of the motion patterns are too hard to discern, for example, fixing hair may be confused with *CellToEar*, the motion of getting an object is identical to that of putting an object, and many actions involve the movement of arms similar to Pointing. 2) Sometimes there are significant cluttered background and cluttered motion background (*e.g.* a crowd of people are moving on the background), which severely degrade the detection performance.

6.3 Evaluation of Sequence-based Performance

After performing the temporal filtering described in Section 5, we obtain the segments of video events. Then, we calculate the detection rate and false alarm defined in Eq. (11) and Eq. (12) to evaluate the sequence-based performance of event detection.

The definition for correct detection is specified by the criterion of event alignment given in the TRECVid 2008 event detection task. The alignment is performed by using the Hungarian algorithm to find the optimal bipartition graph matching in which the system observations (detected and segmented by our approach) are regarded as one set of nodes and the reference observations (labeled from ground truth) are regarded as the second set of nodes in a bipartition graph. Given one system observation O^s and its aligned reference observation O^r, they are matched if and only if

$$Beg(O^r) - \Delta_T \le Mid(O^s) \le End(O^r) + \Delta_T , \qquad (13)$$

where $Beg(\bullet)$, $End(\bullet)$ and $Mid(\bullet)$ represent the beginning, end and midpoint of the event observation's time span, respectively. $\Delta_T = 0.5$ second is a constant differentiating the mappable and un-mappable observations.

Based on the frame-based detection results of the TISR_IG_OF method, we obtain the sequence-based event detection results. Using the TRECVid criterion, the performance is evaluated on all the videos in the development set. The results are listed in Table 3. On average our approach achieves about 8.45% detection rate versus 0.19% false alarm rate for these 3 events. We can see that the performance of our event detection approach is promising on this extremely challenging dataset.

Table 3. Results of sequence-based event detection

Event	# Reference observation	*DR* (%)	*FA* (%)
CellToEar	440	9.60	0.14
ObjectPut	1154	7.34	0.19
Pointing	1671	8.41	0.23
Total/Average	3265	8.45	0.19

As shown in Table 3, the total number of three events is 3265 in the 50-hour videos. The average duration of an event is 34 frames [9]. On the other side, the total number of the video frames in the development set is about 4.75 millions. Apparently, positive event instances are extremely rare compared to the number of negative event instances. Such huge amount of the negative instances demands for very strict SVM classifiers and temporal filtering threshold, otherwise there may be considerable false detections. This is the major reason why the detection rate is low. Moreover, even for the positive event instances labeled by NIST, some of the human figures are too small to provide effective information for the classifier training, which is the second reason for low detection rate. Nevertheless, our approach achieves fairly low false alarm rate. This indicates that our approach can effectively differentiate the actions of interest from such huge amount of negative instances in the video stream.

With the detected event sequences, the TRECVid 2008 event detection task evaluated the submitted systems by the normalized detection cost rate (NDCR) measure [9] which is a weighted linear combination of the missed detection probability and false alarm rate (NDCR = 0 indicates perfect detection performance). According to NIST's notebook papers [9], the top performance system [36], which combines a SIFT based spatio-temporal approach with CNN based and boosting methods, achieved the minimal NDCR = 0.9971, 0.9993, 1.0007 for *CellToEar*, *ObjectPut*, *Pointing*, respectively, on the evaluation set. Adopting the same measure, our method demonstrates competitive performance with the minimal NDCR = 0.9914, 0.9911, and 0.9940 for the three events on our cross-validation set (note the ground truth on the evaluation set is not publically available).

7. CONCLUSIONS AND FUTURE WORK

Human action in videos is an important clue for analysis and understanding of video events. In this paper, we have proposed an effective approach to detect the action-oriented video events in complex scenes based on human action recognition. We introduce a new TISR descriptor for action recognition to capture the patterns in terms of both appearance and motion. Based on the proposed descriptor, the action category is classified by fusing SVM classifiers at multiple spatial pyramid levels. The video events are

segmented using a Gaussian kernel based temporal filtering on the results of frame-based detections.

Our proposed spatio-temporal descriptor is able to effectively encode the characteristics of actions in terms of appearance and motion patterns and is robust to various local variations in complex scenes. By resorting to the bag-of-words technique, the action recognition approach can tolerate spatial and temporal variations of human actions. Compared with the SIFT descriptor which has been extensively applied to object classification and recognition tasks, our TISR descriptor is more powerful for the action-oriented event detection evaluated on the challenging TRECVid 2008 event detection dataset.

The future work includes two directions. First, more kinds of low-level appearance and motion features, *e.g.* Gabor-like filters, will be integrated in the current approach to improve the performance. Second, more sophisticated temporal filtering strategy will be developed to segment the integrated event sequences. The TRECVid event detection task provides a standard benchmark dataset for action-oriented event detection to facilitate fair performance comparison among different algorithms and will surely promote the research of video event detection.

8. REFERENCES

[1] F. Wang, Y.G. Jiang, and C.W. Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. ACM Multimedia*, 2008, pp. 239-248.

[2] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. Int. Conf. Computer Vision*, 2005, vol. 1, pp. 166-173.

[3] D. Xu and S.F Chang, "Visual event recognition in news video using kernel methods with multi-level temporal alignment," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1-8.

[4] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 873-889, 2001.

[5] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing, "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game," in *Porc. ACM Multimedia*, 2006, pp. 431-440.

[6] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *Proc. Int. Conf. Computer Vision*, 2007, pp. 1-8.

[7] 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, pp. 1-8.

[8] C. Schuldt, I. Laptev, and B. Caputa, "Recognizing human actions: a local svm approach," in *Proc. Int. Conf. Pattern Recognition*, 2004, pp. 1-8.

[9] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid. http://www.itl.nist.gov/iad/mig//tests/trecvid/2008/doc/EventDet08-EvalPlan-v07.htm. http://www-nlpir.nist.gov/projects/tvpubs/tv8.slides/event-detection.pdf.

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

[11] H. Buxton, "Learning and understanding dynamic scene activity: a review," *Image and Vision Computing*, vol. 21, pp. 125-136, 2003.

[12] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 34, no. 3, pp. 334-352, 2004.

[13] J. Shen, D. Tao, and X. Li, "Modality mixture projections for semantic video event detection," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1587-1596, 2008.

[14] M. Xu, L. Duan, C. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, 2003, pp. 189-192.

[15] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted American football video by highlight selection," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575-586, 2004.

[16] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and web-casting text," in *Proc. ACM Multimedia*, 2006, pp. 221-230.

[17] L. Xie, P. Xu, S.F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letter*, vol. 25, no. 7, pp. 767-775, 2004.

[18] M.L. Shyu, X. Xie, M. Chen, and S.C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 252-259, 2008.

[19] C.G.M. Snoek and M. Worring, "Multimedia event-based video indexing using time intervals," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 638-647, 2005.

[20] D.A. Sadlier and N.E. Oconnor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225-1233, 2008.

[21] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473-1488, 2008.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Porc. The IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[23] M. Han, W. Xu, H. Tao, and Y. Gong, "An algorithm for multiple object trajectory tracking," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2004, pp. 864-871.

[24] M. Yang, F. Lv, W. Xu, and Y. Gong, "Detection driven adaptive multi-cue integration for multiple human tracking," in *Proc. Int. Conf. Computer Vision*, 2009.

[25] G.A. Korn and T.M. Korn, *Math handbook for scientists and engineers*, New York: McGraw-Hill, 1968.

[26] M. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements and action," *Nature Reviews Neuroscience*, vol. 4, pp. 179-192, 2003.

[27] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. Int. Conf. Computer Vision*, 2007, pp. 1-8.

[28] S. Lazebnik, c. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2169-2178.

[29] V. Vapnik, *The nature of statistical learning theory*, New York: Spinger-Verlag, 1995.

[30] Y.G. Jiang, C.W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. ACM Int. Conf. Image and Video Retrieval*, 2007, pp. 494-501.

[31] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[32] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. Int. Conf. Computer Vision*, vol. 2, 2003, pp. 726-733.

[33] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", in *Advances in Large Margin Classifiers*, Cambridge: MIT Press, 1999.

[34] R. Duda and P. Hart, *Pattern classification and scene analysis*, New York: John Wiley & Sons Inc, 1973.

[35] B.K.P. Horn and B.G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185-203, 1981.

[36] F. Lv, W. Xu, M, Yang, K. Yu, G. Zhu, and Y. Gong, "Surveillance event detection," TRECVid notebook paper in *Proc. TRECVid workshop*, 2008.