Elsevier Editorial System(tm) for Signal Processing
Manuscript Draft

Manuscript Number: SIGPRO-D-11-01271R1

Title: STV-based Video Feature Processing for Action Recognition

Article Type: Special Issue: MMIndexing

Keywords: Spatio-Temporal Volume; 3D segmentation; Region-Intersection; Video event detection

Corresponding Author: Dr Zhijie Xu, PhD

Corresponding Author's Institution: University of Huddersfield

First Author: Jing Wang, PhD, MPhil, BEng

Order of Authors: Jing Wang, PhD, MPhil, BEng; Zhijie Xu

Abstract: Video recordings can provide rich and intuitive information on dynamic events occurred over a period of time such as human actions, crowd behaviours, and other subject pattern changes in comparison to still image-based processes. However, although substantial progresses have been made in the last decade on 2D image processing and its applications such as face matching and object recognition, video-based event detection still remains one of the most difficult challenges in computer vision research due to the wide range of continuous or discrete input signal formats and their often ambiguous analytical features. In this paper, a spatio-temporal volume (STV) and region intersection (RI) based 3D shape-matching method has been proposed to facilitate the definition and recognition of human actions recorded in videos. The distinctive characteristics and the performance gain of the devised approach stemmed from a coefficient factor-boosted 3D region intersection and matching mechanism developed in the programme. This research has also investigated techniques for efficient STV data filtering to reduce the amount of voxels (volumetric-pixels) that need to be processed in each operational cycle in the implemented system prototype. Encouraging features and improvements on operational performance have been registered in the corresponding experiments.

STV-based Video Feature Processing for Action Recognition

Jing Wang, Zhijie Xu*
School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, United Kingdom

Abstract

Video recordings can provide rich and intuitive information on dynamic events occurred over a period of time such as human actions, crowd behaviours, and other subject pattern changes in comparison to still image-based processes. However, although substantial progresses have been made in the last decade on 2D image processing and its applications such as face matching and object recognition, video-based event detection still remains one of the most difficult challenges in computer vision research due to the wide range of continuous or discrete input signal formats and their often ambiguous analytical features. In this paper, a spatio-temporal volume (STV) and region intersection (RI) based 3D shape-matching method has been proposed to facilitate the definition and recognition of human actions recorded in videos. The distinctive characteristics and the performance gain of the devised approach stemmed from a coefficient factor-boosted 3D region intersection and matching mechanism developed in the programme. This research has also investigated techniques for efficient STV data filtering to reduce the amount of voxels (volumetric-pixels) that need to be processed in each operational cycle in the implemented system prototype. Encouraging features and improvements on operational performance have been registered in the corresponding experiments.

Keywords: Spatio-Temporal Volume; 3D segmentation; Region-Intersection; Video event detection

1. Introduction

* Corresponding author: Dr. Zhijie Xu
Email: z.xu@hud.ac.uk
Tel: 0044 1484 472156
Fax: 0044 1484 421106
School of Computing and Engineering, University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom

Computer vision theories and practices have been experiencing an accelerated development period over the last 2 decades. Demands from applications such as intelligent surveillance systems, machine vision, robotics and automatic guided vehicle (AGC), innovative human computer interface (HCI), and even digital entertainment and computer games have pushed this trend. As an important application of video event detection technologies, action recognition is one of the hotly-pursued areas in computer vision research for automatically detecting and interpreting real-world events and activities such as human gestures, crowd actions, or other object patterns through extracting and analysing video features denoted in the spatial and temporal reference frame.

As widely recognized, raw video data often suffer from high noise ratio and low resolution that require tedious and time-consuming processes to clean up at a frame-by-frame (FBF) basis. This research aims at investigating 3D Spatial Temporal Volume based event descriptors for video content analysis and event detection. The rationale behind this research is to develop robust and intuitive video processing methodologies and techniques for tackling challenging tasks such as large video database indexing and querying, automated surveillance system design, and Internet-base online video management.

Generally speaking, the large variations on conditions where videos being made and the complicated nature of human gestures and postures are still posing great challenges to the conventional pixel-based human action recognition strategies. The complexity exposed to video event detection tasks can be classified into three categories. Firstly, the semantic meaning of an “event” in a video is often ambiguous since the variations of potential “event makers” defined by a particular application or system operators. Secondly, due to the limitations of today’s videoing equipment and storage restrictions, real life video data, especially those from surveillance systems and Closed-Circuit Television Systems (CCTV), inherit great technical difficulties to process. For example, the boundaries between an “event” signal and its “background” noise is often either inexplicit or occluded, which renders a complete separation of the two signals almost impossible. In many recent pilot researches, a

background is often simplified as static sections in continuous video frames. However, this presumption is not always applicable in complex real life scenarios, i.e., multiple dynamic objects or illumination changes can all cause confusions and blurriness on that. The third difficulty can be caused by the uncertainty of video event durations. The time-elapsed factor for encapsulating a discrete event is closely coupled with the nature of the event and the videoing sampling rates which might be substantially varied for different application systems.

Compared with the often non-generic local and 2-dimentional (2D) feature-based approaches such as head shape detection, and torso-limb spatial relations, the emerging concept of the spatial and temporal feature space and the so-called Spatio-Temporal Volume (STV) data structure - first introduced by Adelson and Bergen [1] - have shown their promising global feature representation potentials for 3D and dynamic video event feature representation and pattern recognition. A STV model is capable of encapsulating static and dynamic video content features, hence simplifying an event recognition task into corresponding 3D geometric feature extraction and matching operations. As illustrated in Figure 1, the STV is a volume space confined by 3D coordinates system denoted by X, Y (spatial) and T (temporal) axes, in which a STV model can be represented by a 3D shape formed by a stack of 2D arrays of pixels, called voxels standing for volumetric-pixels.

In this research, a spatio-temporal volume (STV) and region intersection (RI) based 3D shape-matching method has been proposed to facilitate the definition and recognition of human action events recorded in videos. An innovative pattern recognition algorithm has been developed to harness the promising characteristics of the STV event models. The core of this approach is built upon the region intersection (RI) method to compare STV “shapes” extracted from video inputs to pre-defined 3D event templates.

It has long been understood that many 2D DIP and pattern recognition algorithms can be readily extended to 3D domains using higher dimensional vectors for defining complex features such as 3D curves, shapes, and volumes. The methodology applied in this research

has started with an optimized 3D over-segmentation operation to detect the boundaries and feature distributions of the studied STV model that ignores the often confusing semantic differences between an event "actor" and its "background". This "background-independent" operation reduces the false-positive rate in differentiating signals from "noises" under complex real-world settings.

After the feature extraction phase, a calibration mechanism is activated to measure the so-called normalized RI distances. This process evaluates the distribution of the segmented regions and then calibrates the matching distance by using a coefficient factor to record the linear corrections required. Compared with other conventional RI-based matching approaches, this extra step further improves the operational robustness at event recognition stage.

This paper will focus on two main contributions from the research to the action recognition domain:

- An innovative segmentation algorithm for extracting voxel-level features has been developed based on a hybrid discontinuity and similarity-based segmentation model through combining Mean Shift (MS) clustering and the graph-based region description method.
- A novel extension of the Region Intersection technique has been realized for human action recognition by using a coefficient factor-boosted template matching method; that is capable of superior performances under complex and real-world videoing conditions.

The rest of this paper is organized in the following order: Section 2 reviews the state-of-the-art for video event detection and STV-based feature processes. Section 3 introduces an innovative 3D STV segmentation strategy developed in this research. Section 4 focuses on the region-based template matching using the event shapes extracted from video inputs and a set of single human action events assembled. Section 5 highlights the implementation strategies for system prototyping. The experiment designs are reported in Section 6. Section 7 concludes the research with discussions on results acquired and planned future works.

2. Literature Review

Video event detection researches encompass a wide spectrum of studies from basic DIP technologies and video processing to pattern analysis and even biological vision systems. After over 30 years of intensive research since its birth, progresses have been made on many fronts with extensive applications found in industry, such as traffic monitoring systems, CCTV-based security and surveillance networks, and robotic control. This section will focus on the prominent works to date on video event definition and STV-based template matching.

2.1. Action Event Definition

Generally speaking, a video event can be classified as single-human-based, such as gestures and postures and multi-human-based like crowd behaviours. For specific research problems or applications, many pilot works have defined specific conditions to simplify the settings of the research platform. A comprehensive survey on those conditions can be found in Moeslund's paper [2] published in 2001. Those restrictions have simplified the abstracted models and reduced the variables needed for dealing with event signals in their feature spaces, but the downside of those approaches is the rigidity of the algorithms and their compromised suitability for general applications.

Single-human-based actions have often been used for tracking individuals, recognizing gestures, and monitoring the change of subject behaviours. Guler et al. [3] have published a paper on tracking individuals who left a baggage behind in an open space. The University of Reading had developed a real-time system for verifying those concepts and techniques [4]. Popular computer vision databases such as KTH [5], Weizmann [6] and Inria XMAS [7] have also encouraged researches into single-human-based action recognition by using 2D/3D models shapes, contours, and stick-based skeleton models. In 2007, Wang and Suter [8] developed an action recognition technique based on human silhouette analysis using the locality preserving projections (LPP) model, which reduces the dimensionality required to transform human actions into lower dimensional spatio-temporal feature space. This development can address problems such as partial occlusion and noisy background.

For multi-human (or crowd) events, particle flows and density models have often been adopted to analyze crowd behaviours with individuals being treated as moving particles. Kilambi and Masoud [9] devised a Kalman filter-based approach for estimating human group sizes. Fleet's [10] dynamic optical flow system and Porikli's [11] trajectory matching algorithm have both deployed the Hidden Markov Model (HMM) for handling the random movements of the monitored crowds. Research progresses on this front have seen proposals and prototypes been developed for applications such as intelligent and adaptive traffic light control, and emergency evacuation systems. In contrast to the single-human-based events, crowd behaviours and events are sometimes difficult to define with accurate semantic interpretations.

This research focuses on investigating proficient global video feature extraction and template matching techniques for STV-based single-human action recognition.

2.2. STV-based Feature Representation

Stemmed from DIP techniques, traditional video event detection approaches relied on spatial or frequency features being extracted from the frame-by-frame-based (FBF) 2D processes [12]. However the FBF-based mechanisms often resulted in the loss of "contextual" information - a main contributor to form and comprehend "dynamic" video events. This draw-back can lead to miss identification and high false-positive rate during an event detection operation, for example, misjudgement of an action (or pose) caused by two persons across over in front of the video camera. One perceived solution for this problem is to construct video capsules which encapsulates both time-tag and frame patterns.

A number of leading computer vision research groups have devised various spatial and temporal frames-based techniques for video analysis, such as flow-based iterations [13], motion history image [14], and local interesting points [15], which focused on the durations and changes of spatial features over time. However, most of these features are gathered from consecutive frames and a small group of pre-determined pixels. The global changes over the

entire scene and the 3D event exercisers - humans or objects - cannot be represented in their entirety. The STV data structure, on the other hand, had long been perceived as a potential solution for the aforementioned problems and to better encapsulating the temporal continuity and "pixel" changes inherited from a video [1].

Constrained by the performance of computer capacities at the time, most of the preliminary researches on STV and its corresponding innovative 3D voxel-based pattern analytical techniques were largely remaining at the level of theoretical discussions. From the middle of 1990s, a number of world-leading research groups in the field attempted to map the STV models to their customized 2D projection planes before applying the conventional pixel-based processing methods for further analysis. One of the representative methods from that era used the so-called "clipping plan" alone the time axis to generate orthographically projected shapes casted from volumetric STV models. These slice-based STV usages have been adopted to infer feature depth information [16], generating dense displacement fields [17], analyzing camera calibration settings [18], categorizing human motion patterns [19], and performing viewpoints synthesis [20] to various degrees of success.

Since the start of the new millennium, the "real" volumetric approaches for 3D-oriented STV processing have been steadily gaining popularity. These new approaches have taken advantages of the volumetric nature of the STV feature space and emphasized on the alteration of shapes, envelopes, and density of the defined feature points in an enclosed space. Research advancements in many related areas such as volume visualization [21] and medical image processing [22] have contributed to the development of these improved voxel-based techniques. Generally speaking, volumetric approaches provide a better vehicle for global representations of the studied subject features. The conceptual model for volumetric video feature processing follows a pipeline [23] that contains several operational phases, including data acquisition, model segmentation, template modelling, and 3D matching.

Based on the review, it is clearly envisaged that the global video feature representation

approaches, such as the STV, will become a valid and popular strategy [23] for video event detection due to its conceptual simplicity and implementation friendliness, which can be regarded as a proximate analogue to the extensively adopted Eigenface model for face reorganization.

2.3. 3D Geometry-based Event Definition

3D geometry-based (henceforth referred as shapes) STV methods treat original STV datasets as "sculpture"-containing (formed by distributed "point clouds") cubes. In 2005, Alper and Mubarak [24] introduced a method to extract 3D human silhouettes from the volumetric space for shape matching. Based on the shape invariants, Alper's method applied the so-called "differential geometric surface properties", such as peaks, pits, valleys and ridges as feature descriptors to denote specific events in the form of vectors in the feature space.

In 2006, through transplanting motion history images onto 3D STV models, Weinland and Ronfard [7] developed a set of view-invariant motion descriptors for human event definition that is capable of representing dynamic events captured in a video by applying Fourier transformations in a cylindrical coordinates system. This process formed a solid foundation for representing view-invariant features in the forms of specific patterns in the frequency domain.

Another 3D shape-invariant analysis method, which was first proposed by Gorelick [6, 25] declared another significant progress in the field. By deploying the Poisson distance equation, the local space-time saliency features and the Hessian-based space-time orientations can be generated to describe volumetric shape features. Based on the aforementioned methods, human gestures can often be classified into a number of types by applying a spectral classification algorithm [26].

As summarized by Shao [27, 28], a recent trend of integrating and enhancing STV global features with a chosen set of local features for calibration has witnessed a degree of success. For example, Jiang *et al*. [29] had introduced an inter-frame constrained local feature

definition method for creating a so-called “convex matching scheme” to facilitate human action detection. Dollár *et al.*’s [30] spatio-temporal cuboid prototyping method had extended the ability of the 2D interest point segmentation technique into Basharat *et al.*’s [31] SIFT-based (Scale Invariant Feature Transform) video retrieval pipeline; Zhai and Shah’s [32] spatio-temporal attention model, Laptev and Lindeberg’s [33] slice-based feature identification techniques, Loper *et al.*’s [34] bag-of-visual-features (BoVF) methods, as well as Shao’s motion and shape feature-based video content segmentation strategies [35] have all been pushing boundaries on this front.

Recently, action recognition researches have been experiencing a weight-shift from the low lever feature and algorithm-oriented development to the semantic-informed machine learning domain. For example, the local feature based Bag of Words (BoW) [36] algorithm and its variations [37] have been used in machine learning-based classification through the sliding window mechanism. Similar efforts, such as Hu’s oriented gradients feature with support vector machines (SVMs) [38] technique and Yang’s motion video patch distance with cascade classification method [39] had partially addressed the occlusion problems often occurred in real-world video recordings.

Based on the spatio-temporal shape feature distribution and the local region grouping, an RI-based distance algorithm harnessing the advantages of both global and local features for recognizing human actions was first introduced by Ke [40]. Although proven effective when segmenting backgrounds of average complexity, the algorithm was less useful when dealing with complex real-world scenes where both extremely large and small textured regions existed alongside each other. This was one of the early motives of this research to investigate system approaches with more consistent performance for real-world applications.

The investigation on STV-based video event detection and system prototyping in this research focuses on two aspects: adaptive feature segmentation optimized shape-based template matching, and rapid-and-lean STV model construction for ensuring runtime system

performance.

3. 3D Hierarchical Feature Extraction

For extracting and representing feature point distributions of a STV model in the event detection pipeline, a 3D hierarchical Pair-wise Region Comparison (PWRC) segmentation algorithm has been developed in this programme and will be referred as e-PWRC throughout this paper with “e” stands for extended.

The baseline 2D PWRC method devised by Felzenszwalb and Huttenlocher [41] is a graph-based clustering technique for representing an image’s pixels and their neighbours as an interconnected graph. The similarity between different regions and the dissimilarity inside a region are compared continuously in the graph. Depending on a “similarity factor” *k*. two regions can be merged into a new and larger region. During the operation, the initially independent regions will keep growing in an iterative fashion until reaching a predefined threshold. Details introduction of this image segmentation approach is beyond the scope of this paper. Readers who interest in its technique and algorithms are suggested to refer [41] for more details.

3.1. 3D STV Segmentation

The PWRC method can be readily extended into 3D domains in theory, for example, the STV space. The devised e-PWRC technique in this research initializes a 3D graph with vertices denoted by volumetric features, for example, using the maximum 26 neighbouring nodes of each voxel to define the edges of the graph. Since each edge can then contain both spatial and temporal information, their alterations and “traces” can naturally reflect the dynamic states of the “tracked” object during the segmentation processes.

However, during the initial feasibility experiments, it was observed that the direct transformation of the 2D PWRC to 3D domain led to several drawbacks. Firstly, the region cluster size *C* cannot be easily adjusted by the factor *k*. As illustrated by a 2D example in Figure 2, larger *k* values lead to larger segmented regions but missed out some important

sections of the object boundaries - the person's left arm. A smaller *k* value ensures all the important object boundaries being identified but in an over-segmentation style with a large amount of overcut and cluttered areas. Secondly, during the region growing stage of the PWRC cycle, the internal difference $Int(C)$ for deciding whether to merge clusters becomes less sensitive, especially when closing to the regions filled with small and random textures since the weights of edges in the graph are computed at per-voxel level, rather than the more representative and efficient regional features. Thirdly, to generate a PWRC graph directly from a raw STV model will introduce huge amount of data to the system platform, for example, a 100×100×100-voxel STV model creates the adjacency of edges on the scale of $(100\times100\times100)^2$ which cannot be handled by any conventional computer programs and runtime memory (RAM), never mention other looping and branching operations on those data.

The proposed solutions for the aforementioned problems were partially motivated by Grundmann's earlier work on video segmentation [42], which focused on the video region description by using texture similarity. It highlighted the importance of similar regions along the temporal domain. To further strengthening the local voxel colour characteristics of the proposed graph-based algorithm, this research has introduced a Mean Shift-based pre-clustering operation for constructing the initial region groups that emphasized the colour features in addition to Grundmann's texture features at the beginning of the e-PWRC operations.

For extracting accurate STV regions from uncontrolled video settings, the e-PWRC deployed an integrated STV manipulation technique consisted of data pre-clustering, histogram-based region description, and hierarchical segmentation, which optimises the performance of factor *k* through the reduction of STV data size and the usage of a region feature-based representation scheme for segmentation.

3.2. e-PWRC Implementation

3.2.1. STV Filtering by Mean-Shift Clustering

To apply PWRC directly to STV segmentation, the vertices in the initialized graph are essentially un-treated voxels. However, large percentage of those voxels only contains static information, which will have no contribution to the eventual feature template matching operations. e-PWRC simplifies the initialized graph using the Mean-Shift (MS) clustering to remove those redundant static features and to combine "similar" voxels into regions before assigning them to the graph vertices based on the similarity of the colour. Compared with the per-voxel-based initialization, the number of vertices of this region-based approach can be reduced significantly. Although still sensitive to noise, the Mean-Shift clustering operation handles small groups of features effectively and can control the segmentation region size flexibly through applying the so-called Parzen window density estimator [43]. In this research, the Mean-Shift clustering for the raw STV is treated as a pre-processing step for data filtering that occur prior to the "primary" STV segmentation step. The core of this data cleansing operation is a Probability Density Function (PDF) ***f(*****x*****)***. In a d-dimensional ($\mathbf{R}^d$) feature space, if $\mathbf{X}$ denotes the collection of feature points with individual point represented as $\mathbf{x_i}, i = 1,2,\ldots n$, the multivariate kernel density estimator with kernel $K_{\mathbf{H}}(\mathbf{x})$ can be computed at the point $\mathbf{x}$ by

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_H(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where H is a symmetric positive $d \times d$ bandwidth matrix.

When applied in 2D image processing, the inputs of the model are often limited to the spatial coordinates and the colour values of 2D pixels in the feature space. In turn, the generated feature space is of 5D (x, y, r, g, b), in which (x, y) denotes the spatial coordinates, and (r, g, b) for the colour of the pixel. These five elements represent a single point x_i in the feature space.

Within the STV domain, the feature space becomes at least 6D defined as (x, y, z, r, g, b),

where (x, y, z) denotes the spatial coordinates, and (r, g, b) for the voxel colours. Therefore, an identical multivariate kernel density estimation process as detailed above can be applied.

It is proven in the e-PWRC feasibility study (see Section 6.3), by carefully selecting the kernel type and window size factor h on colour h_c and location h_l, the regions composed by Mean-Shift clustering can achieve a satisfactory standard with the majority of key features identified.

3.2.2. Histogram-based region description

Post to the pre-clustering process, the initialization of the 3D segmentation graph will be carried out on regional features rather than voxel-level features, hence, significantly reduce the quantity of the required graph vertices. To tackle the feature insensitively problem during region growing stage, this research has introduced tools such as histogram and Histogram Distances to represent region textures in the weighted graph.

After the colour space transformation, high-level region features, such as textures, can be readily defined by the local colour histograms to denote vertices values of a graph. As illustrated in Figure 3, each region contains a normalized local histogram based on the L*a*b* colours [44]. The distribution of these colours in the histogram embodied richer information than the per-voxel features. For example, a texture containing flat distribution of solid colours can be readily represented as multiple peaks in the histogram.

Weights of edges in this new form of graph are defined by the Histogram Distance. In this research, a minimum distance method introduced by Cha et al. in 2002 [45] has been adopted, which can be explained as:

If H_X and H_Y are two histograms that contain n elements each with individual element specified in the style of $h_{X,i}$ and $h_{Y,i}$, where i=1,2,…,n. The distance $D(H_X, H_Y)$ of these two histograms can then been summarised as:

$$D(H_X, H_Y) = \min_{X,Y}\left(\sum_{i,j=1}^{n} d_{\text{nom}}(h_{X,i}, h_{Y,i}), \sum_{i,j=1}^{n} d_{\text{ord}}(h_{X,i}, h_{Y,i}), \sum_{i,j=1}^{n} d_{\text{mod}}(h_{X,i}, h_{Y,i})\right) \quad (2),$$

$$d_{\text{nom}}(x, y) = \begin{cases} 0 & x = y \\ 1 & otherwise \end{cases}$$

$$d_{\text{ord}}(x, y) = |x - y| \quad (3)$$

$$d_{\text{mod}}(x, y) = \begin{cases} |x - y| & |x - y| \le \dfrac{n}{2} \\ n - |x - y| & otherwise \end{cases}$$

Therefore, the weight defined by the minimum distance can be expressed as:

$$w((v_i, v_j)), (i \neq j) = D(Hv_i, Hv_j). \quad (4)$$

3.2.3. Hierarchical Pair-wise Region Comparison

Since region sensitivity of conventional PWRC is only controlled by factor *k*. A fixed *k* value for the region growing operation is insufficient for feature grouping if texture number varies substantially from frame to frame. This drawback has been partially resolved in this research by the e-PWRC's hierarchical segmentation structure and an adaptable and dynamic *k*-selection mechanism.

It is widely recognized that real-world images or video frames especially the outdoor scenes, often contain many large and uniform coloured regions (i.e. blue sky and dark ground) as well as varied textures (i.e. flowers and trees). Most existing segmentation strategies are seemingly specialized in dealing with either prior or latter case. The hierarchical segmentation strategy offers a practical approach to solve this problem by building a pyramid structure for storing and representing raw data, as illustrated in Figure 4.The low resolution images or frames at the top of the pyramid only need to "remember" large coloured blocks with minute details filtered out during the re-sampling operation. The bottom level, on the other hand, records all the details in the original dataset.

The hierarchical segmentation operation starts from the bottom to register all fine details, while the higher level operation builds upon lower level outputs.

As the e-PWRC pushed up to the higher level, small regions are merged, which requires a dynamic mechanism for determining the k value denoted as $k(C)$. Since increasing the k values suppose to "trigger" the merging actions to generate larger regions, hence the $k(C)$ can be defined based on the largest region within the current hierarchical level, as expressed in Equation 5:

$$k(\mathbf{C}_i) = k(\mathbf{C}_{i-1})\left[1 + \frac{2\text{mean}(\mathbf{C}_{i-1})}{\max(\mathbf{C}_{i-1})}\right], \quad (5)$$

where $\mathbf{C}_i$ is the region collection on the i^{th} floor in the hierarchical structure. The current k value is then iterated and updated based on the region size of the lower level which is related to the mean and maximum region sizes. The sensitivity of related parameters were tested in experiment and reported in Section 6.1.

Based on the aforementioned algorithms and techniques, the e-PWRC approach is more adaptable and robust for processing STV models constructed from video inputs. Table 1 illustrates the operational pseudo code for the e-PWRC processes.

In this section, higher-level regional features inherited from raw STV datasets have been refined and clustered using an innovative e-PWRC volumetric segmentation technique. The outputs of this step, including shape boundaries and the distributions of segmented regions describing event models will be used as inputs for pattern analysis and event identification.

4. Template Matching-based Event Detection

Compared with traditional FBF-based video analysis strategies, the significant advantage for detecting subject changes by using the STV model is rooted to its distinctive ability in providing 3D geometric descriptions for dynamic video content features recorded in a footage, which provides a theoretical foundation for event template matching. As illustrated in Section

2.2, video event features can be abstracted and recognized by using either global feature-based shape analysis methods, or local feature-based methods, such as spatio-temporal interest points. A hybrid approach taken both methods' advantages has been developed in this research. This paper will only focus on the global shape-based techniques developed in the programme.

4.1. Over-segmentation for Region Boundary Preservation

As a continuous operational flow in the event-detection system pipeline, the segmentation outputs (highlighted in Section 3) provide shape and boundaries features for representing event profiles in the STV space, which transforms the event recognition operation into a 3D shape matching operation. Various pattern matching techniques, such as Gorelick [25], Alper [24] and Weinland [7], analyze the distributions of boundary segments directly based on the assumption that the segmentation outputs contain "perfectly sorted boundaries". In reality, video events are inherently difficult to be cleanly separated from uncontrolled backgrounds in real world recordings, a condition which was often ignored by fine-tuned laboratory experiments. Many "fake" regions can be falsely identified as interested regions. For example, as illustrated in the Figure 5, the torso of the "actor" in the video has been segmented into many small parts due to the texture variance of his clothes, which is far from "perfect". These small regions caused by over-segmentation are commonly treated as problematic and considered the main cause to the low efficiency of the relevant pattern analysis algorithms due to extra filtering required to "clean" the region boundaries.

In this research, the over-segmented event volumes are not viewed as "further improvement required" but an intermediate output that can be directly fed into the innovative shape-based matching algorithm developed in the programme. A close examination of the Figure 5 reveals that the over-segmentation has effectively identified all the intersected spaces actually contains all the shape boundary fragments (sub-boundaries) in the volume. The matching operations can be carried out based on these sub-boundaries through analyzing their distributions. This approach can be classified into the so-called "Region Intersection" (RI)

template matching category. One of the distinctive features of RI methods is their ability to perform shape-based event detection in challenging real-world settings where event signals are often immersed under complex background noises. The method devised in this research has explored the following design theorem.

4.2. Baseline RI Method

Based on the Set Theory, the mechanism of a baseline RI method can be simplified into 4 typified cases as illustrated in Figure 6, where the cuboid represents an over-segmented STV. An event shape template, highlighted by the bold boundary will shift through the entire volume at runtime to identify any intersections by comparing with the sub-regions. The RI matching algorithm then calculates the sum of all the intersected parts and their matching distances based on the 4 possible scenarios as illustrated in Figure 6.(b) to(e).

These different scenarios can be summarized as:

$$I(E,S_w)=\begin{cases}|E\cap S_w| & if \quad |E\cap S_w|<|S_w|/2\\ |S_w-E\cap S_w| & otherwise\end{cases}, \tag{6}$$

where I denotes the distance, E represents event templates and S_w marks one sub-region during the matching process. The overall distance between an event template and the detected pattern can be written as:

$$I_l(E,S)=\frac{1}{N(E)}\sum_{i=1}^{n}I_l(E,S_w); \tag{7}$$

$$N(E)=\begin{cases}\sum_{i=1}^{n}\left(\frac{|E|}{2}-\frac{|E|}{2^{|E|+1}}C_{|E|}^{|E|/2}\right), & |E|=even\\ \sum_{i=1}^{n}\left(\frac{|E|}{2}-\frac{|E|}{2^{|E|}}C_{(|E|-1)}^{(|E|-1)/2}\right), & |E|=odd\end{cases}, \tag{8}$$

where l denotes the current location of the sliding template window. n and i denote the number of sub-regions and current sub-region, respectively. The $N(E)$ is the normalized factor associated with the distribution of the event template, which can be calculated independently and used as look up table (LUT) during the matching operations.

After scanning through the entire volume in the searching window style, the RI operation will denote all the "matched" locations judged by matching distances less than a pre-defined threshold.

4.3. Coefficient Factor Enhanced RI Distance

The RI baseline method introduced above can detect most event corresponding shapes in a STV. However, the accuracy of this method can be further improved to satisfy real-world settings, where the proposed coefficient factor (CF) serves as a gauging mechanism to verify the accuracy of the computed RI distances. As discussed in Section 3.3.3, real-world video inputs usually contain both large solid coloured areas and detailed textures. The e-PWRC segmentation method can classify these contents into a hierarchical and over-segmented style to handle both large and small sub-regions. But as shown in Figure 7.(d), some extremely small regions around the event shape boundaries can produce substantial normalized RI distances, which is a potential cause for the so-called "false negative" detection results. In the new approach, a reward-and-punish scheme has been established to assign more importance to the larger sub-regions that effectively reduce the distance values. For the smaller regions, contrary measures will be taken automatically.

This evaluation scheme is automatically generated once the STV is imported and is based on a quantifying process on the intersected regions' local histograms to record the size and the numbers of the intersected sub-regions as shown in Figure 7.

The local histograms discussed above are used to compare with the histograms extracted from the controlled and ideal RI matching scenarios (ground truth). When the event actors and backgrounds are perfectly segmented as illustrated in Figure 7.(a), all feature points on the contour of the template will be matched to the segmented patterns, therefore the histogram will show a straight line lying on the horizontal axis.

In real world scenarios, there are three main representative situations when calculating the coefficient factors, as illustrated in Figure 7.(b) - (d), where the event templates are denoted

by the artificial ellipses. In Figure 7.(b), the intersection parts are mainly composed by a large quantity of small sub-regions. The distribution histogram illustrated at the right hand side shows a single peak near the original point. In addition, 7.(b) also indicates template can match arbitrary small sub-regions without normalization (Equation 7). It is worth noting that the CF factor "amplifies" the difference in this case. On the contrast, in Figure 7.(c), the histogram is showing a largely flat curve with small fluctuations indicating a fewer but larger intersected region blocks, which is an ideal situation for the RI matching and should be "reward" after normalization by reducing the difference between the template and pattern. Figure 7.(d) contains both large and small intersectional parts, where the smaller regions are in dominance; therefore the diagram shows a prominent peak in the histogram with smaller variations on other places. Through using the histograms, the distribution of different types of intersectional groups can be evaluated using the normalized χ^2 distance between the current histogram and its ground truth counterpart.

In general, the coefficient factor (CF) can be expressed as a linear transformation from the histogram distance as denoted in Equation 9:

$$\tilde{I}_l(E,S) = I_l(E,S) \cdot [a + b \cdot f_l(E,S)], \tag{9}$$

where $f_l(E,S)$ is the normalized χ^2 distance of the histograms. The lower limited a and slope b control the degree of the correction of the RI distance. The value of the coefficient factor should be around 1, which is the threshold in switching between "rewarding" and "punishing". In the experiments, the range of changes is in between 0.6 and 1.4, which is proven suitable for most of the video datasets tested.

5. Implementation Strategy and System Prototyping

5.1. System Modularization and Data Pre-processing

From the viewpoint of process modularization, computer vision application systems often follow a process pipeline composed of three modules, data acquisition, feature extraction, and pattern recognition. Each module can also be further divided into detailed operations

depending on the deployed processing strategy and algorithms. For example, data acquisition is often encompassing from the stage of receiving sensory signals to data pre-processing like filtering and decompressing.

Original STV models are often extremely large in data sizes since their volumetric 3D nature and rich per-voxel characteristics. This factor often hampered the effort in adopting the STV-based methods for real-world applications, especially for those time-critique video indexing and retrieval applications. In this research, the runtime optimization technique deployed at the data acquisition stage is based on an "on-fly" computer memory management strategy called "volume buffering".

As illustrated in Figure 8.(a), at runtime, the system starts with building a buffer (assigning memory) to the incoming video stream. The index of the first fame and the size of a STV model are customizable and dependant on the pre-defined action template sizes. Once the last event matching step is completed, the buffer assigned for holding the STV model will be freed from the memory to avoid accumulating memory footprints for the next cycle.

The benefit of this design is achieved through harnessing an efficient computer data structure - queue - that enables a first-come-first-serve process order and its intrinsic flexibility in handling arbitrary sizes of data packets. Figure 8.(b) illustrates the STV construction and registration procedures followed in this programme using 1D vectors as an example. Currently, the assignment of the starting frame's index has been simplified by halving the previous STV chunk and using its "middle" frame as the beginning for the next model. A more robust sampling approach, for example a one-third start from the previous STV or even an arbitrary starter, will see a more adaptable reconstruction process with an improved chance to encapsulate an event occurred, but the effect is envisaged as limited with prolonged latency.

The improved RI matching process can be summarized in the pseudo code shown in Table 2, where the settings of the STV size and its starting frame struck index have been parameterized.

5.2. Pipeline Construction and Prototyping

The event detection system prototyped in this programme has developed a process pipeline as illustrated in Figure 9. The functional modules (enclosed by solid-rectangles) denote the methodologies devised and utilized in the system. The modules marked by dashed-rectangles represent system optimisation techniques implemented. The system begins with a video signal acquisition module that generates STV models in the volume buffer. The e-PWRC segmentation then takes place on the models for extracting event shapes prior to the improved RI template matching operations. Both online public video databases, such as KTH [5] and Weizmann [6], and self-made video clips have been used in creating the benchmark human actions, such as “waving hands”, “walking” and “jumping”.

6. Test and Evaluations

A series of experiments have been designed and carried out for validating the system design and benchmarking its runtime performances under both the controlled and real-world settings. The system software was developed using OpenCV 2.2 library functions and the LabVIEW simulation toolkit. The hardware platform included a host PC with an AMD 2.62 GHz Athlon CPU and 2G RAM.

6.1. System Performances on Controlled Datasets

A series of experiments have been carried out in this research to validate the theoretical design of the system elaborated in Section 3 and 4.

The experiments started from processing and recognizing events recorded in the popular KTH computer vision database. Table 3 lists the values of relevant parameters used in the experiment. The event templates applied for RI matching in the KTH dataset were calibrated by averaging the ground truth segmentation results from four volumetric contours in each defined event category.

Figure 10 shows the test results on the detection accuracy based on the confusion matrix acquired from each dataset. The average accuracy of the developed system is 82.0%, which is

slightly better than other popular techniques as listed in Table 4. As illustrated in the confusion matrix, certain events such as “jog-and-run” and “box-and-clap” are difficult to distinguish due to their silhouette similarities and small variation along the temporal axis. One possible solution to such a problem is to combine the machine learning algorithms with the local spatio-temporal features for differentiating human gestures.

Compared with some recent published methods, such as [37, 46], whose accuracies registered on the KTH datasets often reached near 90%, the proposed method seems produced a slightly lower detection rate. This is because the experiment does not involve scale-independent operations during the STV shape matching for simplifying experiment design and highlighting more generic application scenarios for recognizing everyday human activities without strong assumptions and pre-conditions on the settings of backgrounds scenes and human models, which offers improved consistency and robustness for real-world implementation.

An experiment has been carried out for evaluating the sensitivity of the four parameters utilised in the devised system approach, the e-PWRC factor $k(\mathbf{C}_0)$, the hierarchical level n, and the linear factors of the RI coefficients a and b. The same templates employed in the previous accuracy evaluations from each action event categories have been utilised. The independent contributions from each of the 4 parameters have been evaluated by fixing the other three at their optimized value. Figure 11 denotes the relationships between each parameter and the average detection accuracy.

It is noticeable in Figure 11, the curve patterns of the e-PWRC-related parameters $k(\mathbf{C}_0)$ and n have shown similar distributions while the e-PWRC transmitting from over- to under-segmentation. Hence, the system performance can be improved significantly if more accurate over-segmented sub-regions can be extracted. It is also worth noting that the purposed method can still successfully detect about 42% of the queried events even under extremely over-segmented conditions.

The second line of Figure 11 listed the learning parameters *a* and *b* of coefficient factor used in the RI matching. As mentioned in Section 4.3, the value of the coefficient factor should be around 1 based on the normalized χ^2 distance of the histograms, which is the threshold in switching between "rewarding" and "punishing". In the figure 11, the accuracy performance is above 90% when appropriate linear transformation can be applied for the threshold.

As discussed in Section 4.3, the values of the coefficient factor *a* and *b* should be around 1 based on the normalized χ^2 distance of the histograms, which has been set as the threshold in this design for switching between "rewarding" and "punishing". In Figure 11, the accuracy performance is above 90% for detecting the "waving" when appropriate *a* and *b* have been applied to the threshold.

6.2. Performances on uncontrolled datasets

If the aforementioned tests were trials for proof-of-idea, the experiments designed and carried out from this section and onwards were focusing on system performances for real applications with the raw video inputs subjecting to uncontrolled noise level and challenging real application conditions. The parameters adopted for these tests are defined in Table 5.

Since the online Weizmann computer vision database [6] possesses a set of video files containing simple backgrounds and a single event actor in each video clip that was considered ideal for extracting standard and benchmarking event templates. In this research, the entire Weizmann video database has been examined and utilized for template composition through averaging the extracted STV event shapes from the same event category.

A large set of footages recorded in different locations at different time of the day in a university campus have been used in the experiment (see Figure 12), which contains a mixture of action events for comparing with predefined templates. The length of each video is about 600 seconds.

The average time consumption for the event detection processes is around 30 seconds based

on the parameters listed in Table 5, approximately 10 seconds for video segmentation and 20 seconds for template matching.

Figure 13 illustrates the Receiver Operating Characteristic (ROC) and Recall-Precision (RP) curves generated from the region intersection experiments recorded at a 10% incremental step size on the threshold. It is evident that the proposed system in this programme produces superior performances against other benchmarked systems attributing to the introduction of the coefficient factor verifying mechanism as explained in Section 4.3. In addition, the innovative Hierarchical e-PWRC method facilitated the abstraction of more accurate shape features.

For benchmarking the proposed approach, the Recall-and-Precision experiments have been carried out on the widely used video datasets containing complex and dynamic backgrounds [47]. The dataset contains 48 videos and 110 distinguishable human actions labelled as one-hand waving, two-hand waving, picking, pushing, and jumping-jacks, etc. Figure 14 shows the detecting accuracies denoted by relevant RP curves. The proposed method and its corresponding techniques have shown more consistent and improved performance in comparisons with the parts-based template matching and flow correlation method [40].

For testing the prototype system's performance on visual tagging and video indexing, a set of real-world CCTV files have been utilized. The original video clips were downloaded from the social website – YouTube [48]. Figure 15 shows snapshots and detection result from the videos showing a number of pedestrians tripping and falling down at a spot near the entrance of a building. The developed system has been used to detect and denote all the "tripping-and-falling" events from the footages. The average length of the input videos is approximately 5 minutes with 8 "tripping-and-falling" events. The first 4 events are used for defining the template. All the remaining 4 events were successfully identified and denoted. It is clearly visible from the snapshots that the original videos were filled with noise signals and containing moving objects such as vehicles and pedestrians in the scenes.

6.3. MS Clustering Performance

The MS clustering performance for the initialization of the e-PWRC graph regions is also tested in the experiment by using the campus dataset. The proposed method applies the "multivariate normal kernel" [49] for the MS clustering. Relative small bandwidth was chosen for keeping most important image.

It is evident in the experiments that the e-PWRC method has been benefitted significantly from its MS-based hierarchical design and the dynamic parameter controls. Similar conclusions were reached by comparing the e-PWRC-empowered method with the Coefficient Factor (CF)-boosted RI method as shown in Figure 16. The event detection accuracy has registered a 9% increase by using the devised e-PWRC framework.

Different from Grundmann's approach [42], the vertices of the initial PWRC graph and $k(\mathbf{C}_0)$ in the Equation 5 in the new method were based on the aforementioned MS clustering results rather than the direct employment of the original colours of each voxel. The improved $k(\mathbf{C}_0)$ corresponds to the MS parameters, H_c and H_l, to highlight the importance of the low-level colour features during the initial PWRC merging calculations at the bottom layer within the hierarchical structure. After the first merging operation is completed, the PWRC segmentation will literally shift from a colour-based process to texture-based calculations. The experiment results have shown that during the feature extraction stage, the MS pre-clustering actually serves as an effective measure for simplifying the initial graph setting.

As indicated by the experimental results, the e-PWRC method has proven its effectiveness for providing quality outputs for the following RI matching operations, especially when subjecting to complex real-world conditions. Accompanied by the optimization measures employed at the template matching stages, the process pipeline and its various operational models have shown satisfactory consistency and robustness.

6.4. RI Matching Performance and Template Representativeness

In the system prototype design, templates are formed by averaging event boundaries from the

same event groups. In this experiment, the matching accuracy was further evaluated by taking into account of template representativeness that is the number of "similar" events used for producing the averaged template. The template samples were all selected from Weizmann dataset and the self-recorded campus videos were used for assessment. The final accuracy performances of each event categories are illustrated in Figure 17.

Experiment results shown in Figure 17 indicate that the accuracy improvements can be significant when more representative templates were deployed. However, it was also observed that if too many samples - even with just small distinctions - were used, the final "averaged" template can carry substantial "fuzziness" that can lead to miss-identification to other event types. For this project, a sample template number of 7 seemingly struck the balance between the template representativeness and distinctiveness.

7. Conclusions

This research programme has developed a STV-based strategy and related techniques to tackle the challenging problems faced by the researchers and applications in video event detection domain. An innovative image segmentation technique has been developed in this research during the feature extraction phase of the operational pipeline. The proposed extended Pair-wise Region Comparison (e-PWRC) method is a set of hierarchical segmentation operations for classifying STV regions based on regional colour and texture features. Its baseline algorithm follows an iterative mechanism and updates each cluster in every cycle by comparing their inner difference and similarities with neighbouring clusters. In this research, a graph-oriented comparison approach has been applied into the STV space. Based on the theoretical study and practical trials, the e-PWRC segmentation strategy developed in this project has proven its effectiveness and efficiency when extended from 2D to 3D feature spaces.

Based on the early works from Ke et al.'s [40], an improved Region Intersection (RI) method has been developed for recognizing video events by matching the global features assembled

from the over-segmented e-PWRC outputs with event templates. The matching outputs can be further refined by deploying an evaluation scheme based on the so-called coefficient factors through verifying the RI distances. These devised recognition procedures have shown their distinctive advantages when dealing with real-world video inputs. To maintain runtime performance of the research system, the process pipeline and system prototype have adopted a modularised design and were enhanced by a number of optimization techniques, such as volume buffering for pre-filtering the STV models.

The research system of its current form can only handle video events that possess distinctive convex-style shape changes. It cannot register motions occurred "inside" of a volume (concave), for example, a front view of a human hand-clapping event. It is envisaged that the future works will focus on integrating other analytical features in the volume space to yield more intrinsic and non-intuitive information for interpretation.

Reference

[1] E.H. Adelson, J.R. Bergen, Spatiotemporal Energy Models for the Perception of Motion, Journal of the Optical Society of America A. 2 (1985) 284-299.

[2] T.B. Moeslund, E. Granum, A Survey of Computer Vision-based Human Motion Capture, Computer Vision and Image Understanding. 81 (2001) 231-268.

[3] S. Guler, J.A. Silverstein, I.H. Pushee, Stationary Objects in Multiple Object Tracking, in: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), 2007, pp. 248-253.

[4] L. Patino, F. Bremond, M. Evans, A. Shahrokni, J. Ferryman, Video Activity Extraction and Reporting with Incremental Unsupervised Learning, in: Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010, pp. 511-518.

[5] C. Schuldt, I. Laptev, B. Caputo, Recognising Human Actions: A Local SVM Approach, in: International Conference on Pattern Recognition, Cambridge, UK, 2004, pp. 32-36.

[6] L. Gorelick, M. Galun, E. Sharon, R. Basri, A. Brandt, Shape Representation and Classification Using the Poisson Equation, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1991-2005.

[7] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding. 104 (2006) 249-257.

[8] L. Wang, D. Suter, Learning and Matching of Dynamic Shape Manifolds for Human Action Recognition, IEEE Tansactions on Image Processing. 16 (2007) 1646-1661.

[9] P. Kilambi, O. Masoud, N. Papanikolopoulos, Crowd Analysis at Mass Transit Sites, in: Intelligent Transportation Systems Conference, Toronto, Ont., 2006, pp. 753-758.

[10] D.J. Fleet, M.J. Black, Y. Yacoob, A.D. Jepson, Design and Use of Linear Models for Image Motion Analysis, International Journal of Computer Vision. 36 (2000) 171-193.

[11] F. Porikli, Learning Object Trajectory Patterns by Spectral Clustering, in: IEEE International Conference on Multimedia and Expo, Taipei, 2004, pp. 1171-1174.

[12] D.M. Gavrila, The Visual Analysis of Human Movement: A Survey, Computer Vision and Image Understanding. 73 (1999) 82-98.

[13] S.S. Beauchemin, J.L. Barron, The Computation of Optical Flow, ACM Comput. Surv. 27 (1995) 433-466.

[14] A.F. Bobick, J.W. Davis, The Recognition of Human Movement using Temporal templates, Pattern Analysis and Machine Intelligence, IEEE Transactions on. 23 (2001) 257-267.

[15] J. Shi, C. Tomasi, Good Features to Track, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 1994, pp. 593-600.

[16] H.H. Baker, R.C. Bolles, Generalizing Epipolar-Plane Image Analysis on the spatiotemporal surface,

International Journal of Computer Vision. 3 (1989) 33-49.
[17] Y. Li, C.K. Tang, H.Y. Shum, Efficient Dense Depth Estimation from Dense Multi-Perspective Panoramas, in: Eighth IEEE International Conference on Computer Vision, Vancouver BC, 2001, pp. 119-126.
[18] G. Kuhne, S. Richter, M. Beier, Motion-based segmentation and contour-based classification of video objects, in: Proceedings of the ninth ACM international conference on Multimedia, ACM, Ottawa, Canada, 2001.
[19] C.W. Ngo, T.C. Pong, H.J. Zhang, Motion Analysis and Segmentation through Spatio-Temporal Slice Processing, IEEE Transactions on Image Pricessing. 12 (2003) 341-355.
[20] A. Rav-Acha, S. Peleg, A Unified Approach for Motion Analysis and View Synthesis, in: 2nd International Symposium on 3D Data Processing, Visualization and transmission, 2004, pp. 717-724.
[21] M.M. Yeung, Y. Boon-Lock, Video visualization for compact presentation and fast browsing of pictorial content, Circuits and Systems for Video Technology, IEEE Transactions on. 7 (1997) 771-785.
[22] H. Peng, Z. Ruan, F. Long, J.H. Simpson, E.W. Myers, V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets, Nature Biotechnology. 28 (2010) 348-353.
[23] R. Popper, A Survey on Vision-based Human Action Recognition, Image and Vision Cimputing. 28 (2010) 976-990.
[24] Y. Alper, S. Mubarak, Actions Sketch: A Novel Action Representation, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 984-989 vol. 981.
[25] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as Space-Time Shapes, Pattern Analysis and Machine Intelligence, IEEE Transactions on. 29 (2007) 2247-2253.
[26] J. Tangelder, R. Veltkamp, A survey of content based 3D shape retrieval methods, Multimedia Tools and Applications. 39 (2008) 441-471.
[27] L. Shao, R. Mattivi, Feature Detector and Descriptor Evaluation in Human Action Recognition, in: ACM International Conference on Image and Video Retrieval 2010 (CIVR 2010), Xi'an, China, 2010, pp. 477-484.
[28] L. Shao, X. Chen, Histogram of Body Poses and Spectral Regression Discriminant Analysis for Human Action Categorization, in: F. Labrosse, R. Zwiggelaar, Y. Liu, B. Tiddeman (Eds.) British Machine Vision Conference 2010, BMVA Press, Aberystwyth, UK, 2010, pp. 88.81-88.11.
[29] H. Jiang, M.S. Drew, Z.N. Li, Successive Convex Matching for Action Detection, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, , 2006, pp. 1646-1653.
[30] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, 2005, pp. 65-72.
[31] A. Basharat, Y. Zhai, M. Shah, Content Based Video Matching using Spatiotemporal Volumes, Computer Vision and Image Understanding. 110 (2008) 360-377.
[32] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in: Proceedings of the 14th annual ACM international conference on Multimedia, ACM, Santa Barbara, CA, USA, 2006.
[33] I. Laptev, T. Lindeberg, Space-time Interest Points, in: 9th IEEE International Conference onComputer Vision, 2003. Proceedings., 2003, pp. 432-439 vol.431.
[34] A.P.B. Lopes, R.S. Sloveira, J.M.d. Almeidq, A.d.A. Araujo, Spatio-Temporal Frames in a Bag-of-visual-features Approach for Human Actions Recognition, in: 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, Rio de Janiero, 2009, pp. 315-321.
[35] L. Shao, L. Ji, Y. Liu, J. Zhang, Human Action Segmentation and Recognition via Motion and Shape Analysis, Pattern Recognition Letters. 33 (2012) 438-445.
[36] J.C. Niebles, H. Wang, F. Li, Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, International Journal of Computer Vision. 79 (2008) 299-318.
[37] J. Yuan, Z. Liu, Y. Wu, Discriminative Subvolume Search for Efficient Action Detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2442-2449.
[38] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, T.S. Huang, Action Detection in Complex Scenes with Spatial and Temporal Ambiguities, in: IEEE 12th International Conference on Computer Vision (ICCV), 2009.
[39] W. Yang, Y. Wang, G. Mori, Efficient Human Action Detection using a Transferable Distance Function, Lecture Notes In Computer Science; Computer Vision - ACCV 2009. 5995/2010 (2009) 417-426.
[40] Y. Ke, R. Sukthankar, M. Hebert, Volumetric Features for Video Event Detection International Journal of Computer Vision. 88 (2010) 339-362.
[41] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient Graph-Based Image Segmentation, International Journal of Computer Vision. 59 (2004) 167-181.
[42] M. Grundmann, V. Kwatra, M. Han, I. Essa, Efficient Hierarchical Graph-Based Video Segmentation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2010, San Francisco, USA, 2010.
[43] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, Hoboken, 1973.
[44] B. MacEvoy, CIELUV uniform color space. http://www.handprint.com/HP/WCL/color7.html#CIELUV, in, Handprint, 2010.
[45] S.H. Cha, S.N. Srihari, On measuring the distance between histograms, Pattern Recognition. 35 (2002) 1355-1370.
[46] C. Liangliang, L. Zicheng, T.S. Huang, Cross-dataset action detection, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1998-2005.
[47] K. Yan, R. Sukthankar, M. Hebert, Event Detection in Crowded Videos, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1-8.
[48] People Falling Down Funny Invisible Curb Slow Motion, in, YouTube, 2012.

[49] D. Comaniciu, Mean Shift: A Robust Approach Toward Feature Space Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence. 24 (2002) 603-619.

Figure Captions:

Figure 1	A "falling down" video event defined by a STV shape.
Figure 2	Factor k-control for dealing with noisy video signals. Since the factor dominates PWRC segmentation performance, a dynamic k value based on the scene complicity is required for the flexibility improvement under uncontrolled real-world settings.
Figure 3	Graph formed by using region histogram. The histograms represent the local texture regions initialized by MS clustering.
Figure 4	e-PWRC Hierarchical graph representation.
Figure 5	Video segmentation result by using hierarchical e-PWRC. The shape, size and distribution of the over-segmented regions can be used for STV-based event matching since the regions actually contain all the shape boundary fragments in the volume. In this research, the shape and topological distribution of the regions are matched by using region intersection method. The size and its statistical distributions, defined as a matching coefficient, are evaluated based on the histogram of intersected regions.
Figure 6	RI template matching algorithm and the four typical scenarios.
Figure 7	The coefficient factor designed for RI distance. (a): Template ground truth and size-quantity-histogram of the intersection regions, which is a straight line lying on the horizontal axis. (b)-(d): Three over-segmented region distributions during the intersection. The left column shows diagrams illustrating the RI distance (shade areas). The templates are denoted by red boundaries and possible patterns are highlighted by bold area. The snapshots in the middle indicate each distribution under a real-world scenario highlighted in red ellipses area (artificial templates). The right histograms are size-quantity statistic.
Figure 8	Volume buffer structure for large video indexing. (a): Volume buffer data structure. (b): Volume buffering procedures.
Figure 9	System pipeline.
Figure 10	The KTH confusion matrix.
Figure 11	Parameter sensitivity test result
Figure 12	Campus Dataset.
Figure 13	System performance on campus datasets. The experiment also indicated the accuracy improvement based on the proposed method.
Figure 14	Benchmarking Recall-and-Precision experiments on complex video datasets. System performances on crowded video dataset compared with benchmarked method.
Figure 15	Detecting and indexing exercise of the"Falling down" events. The templates are highlighted in the first row. Selected snapshots from video feed-ins are lined up in the second row, where the ground truths are denoted by the red overlaying areas, and the indexed results highlighted by the orange regions.
Figure 16	The confusion matrices on the campus dataset for evaluating the independent MS contribution within e-PWRC.
Figure 17	Accuracy impact from averaged template numbers.

Figure 1

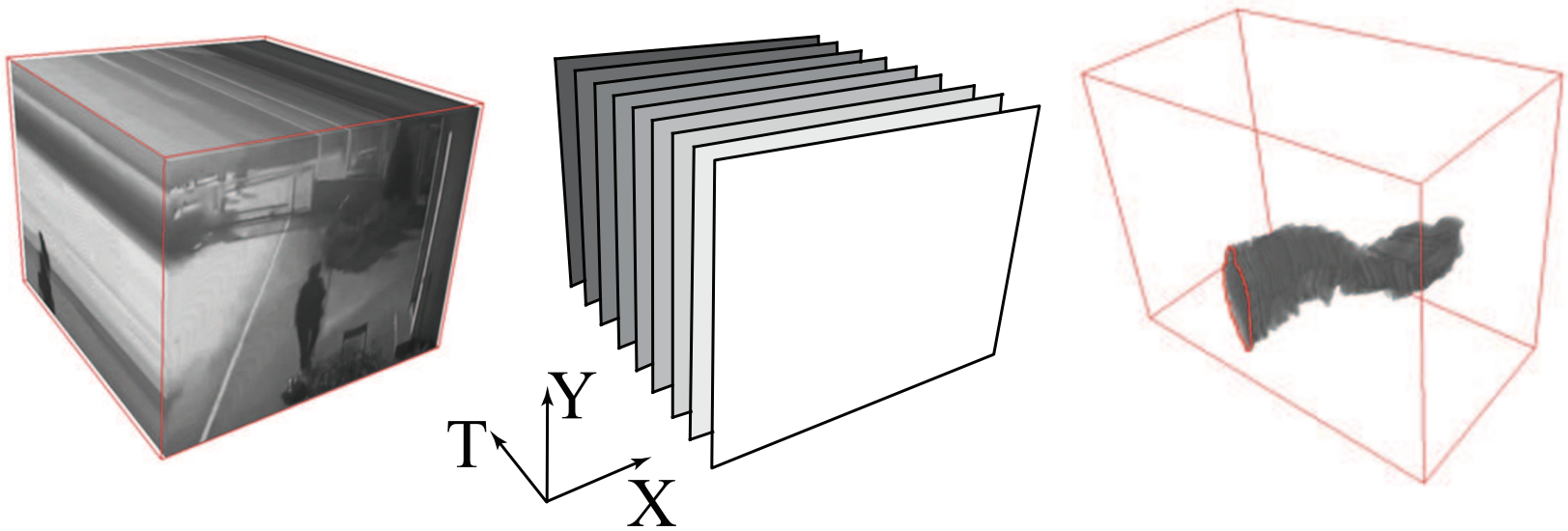


Figure 2

Figure 3

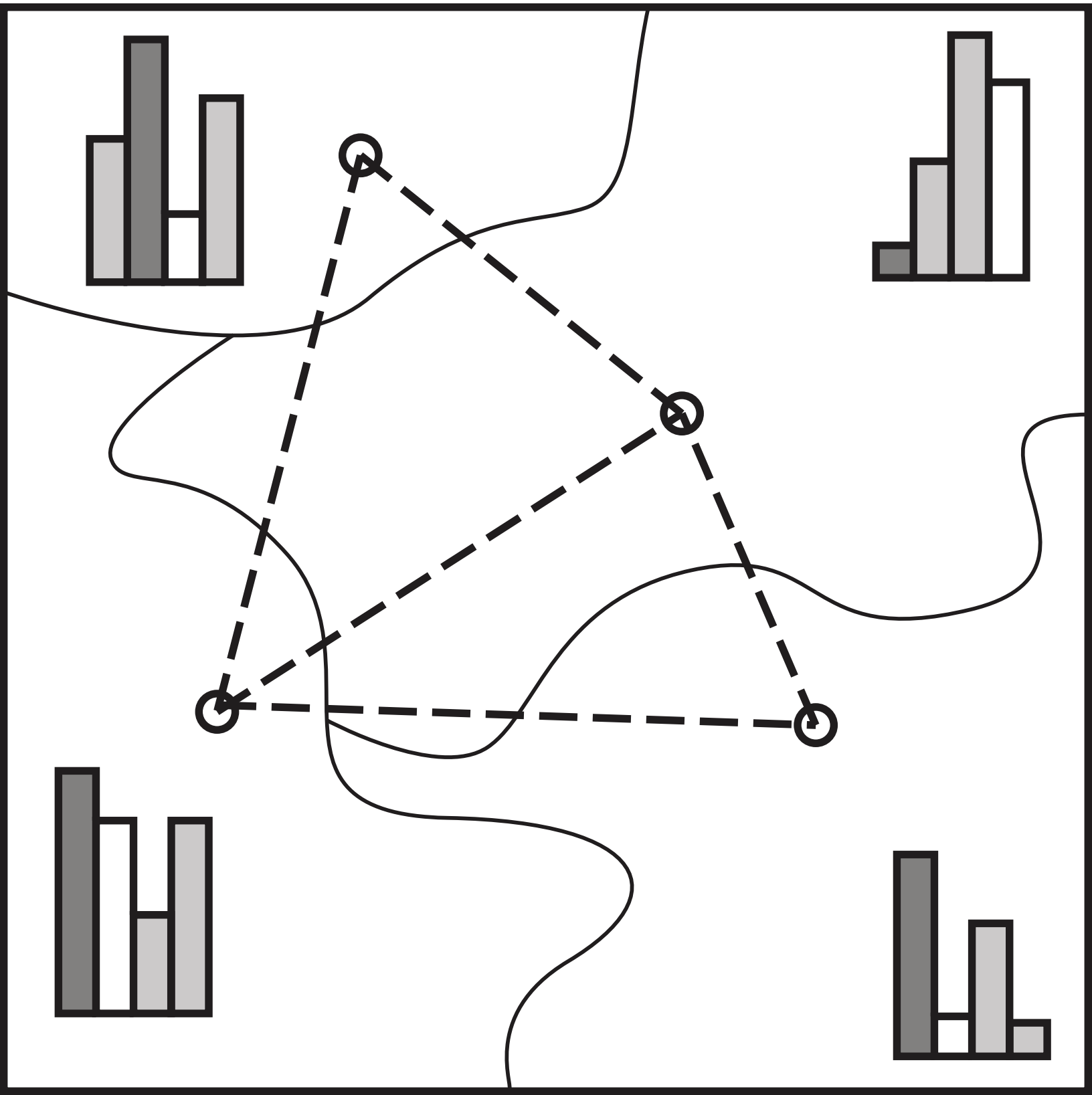
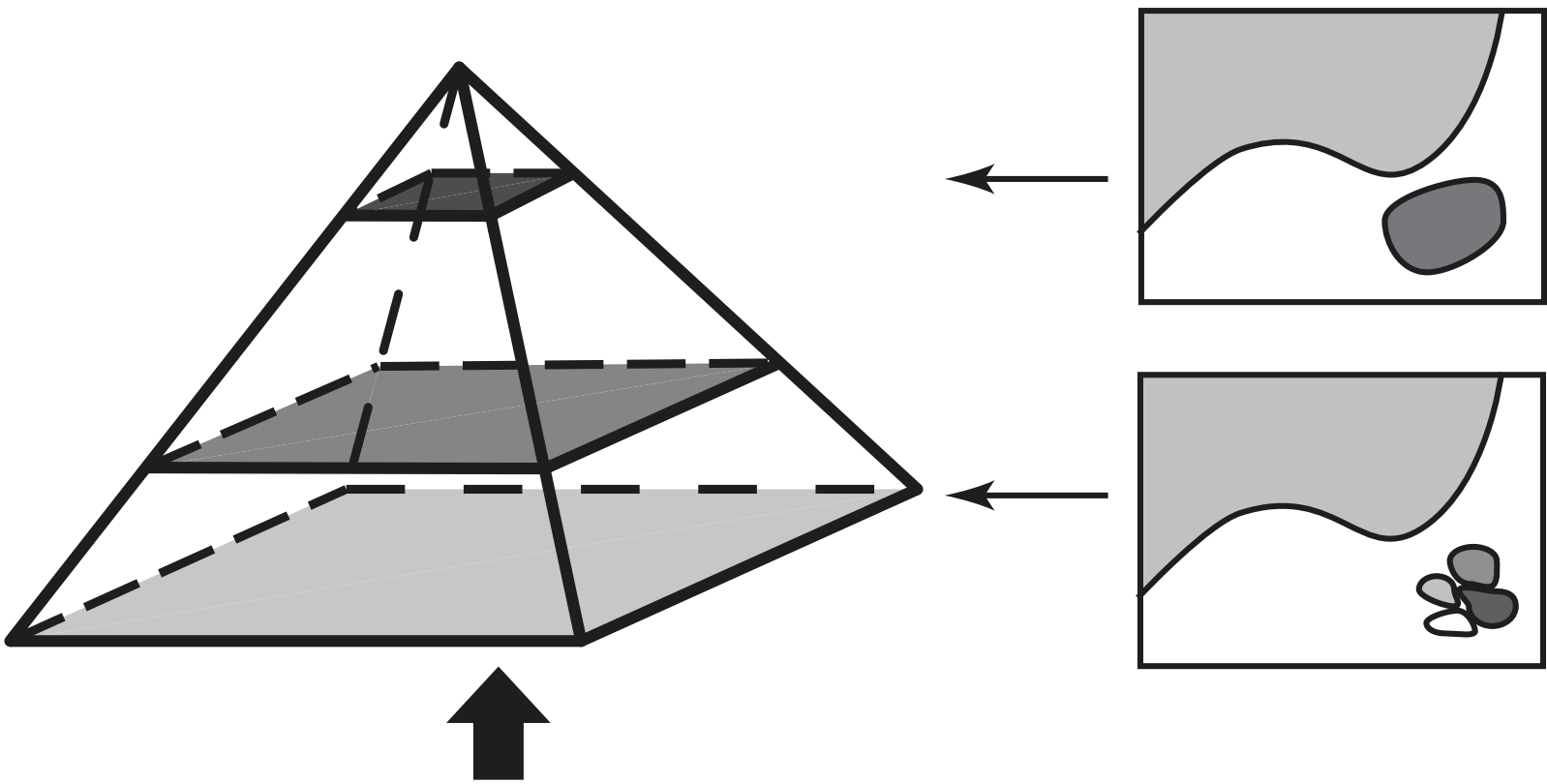


Figure 4



Features from MS-based Pre-segmentation

Figure 5

Click here to download high resolution image

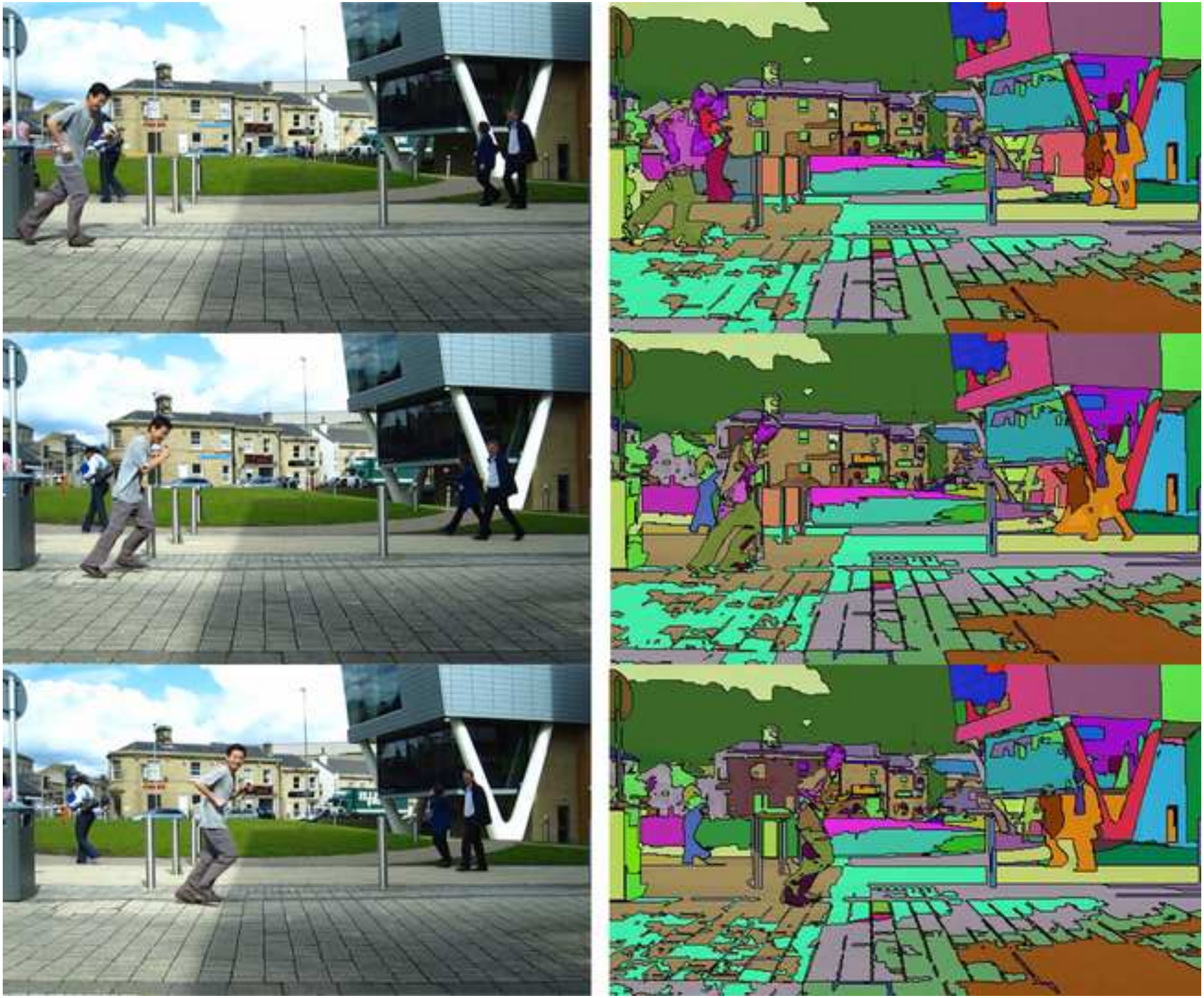
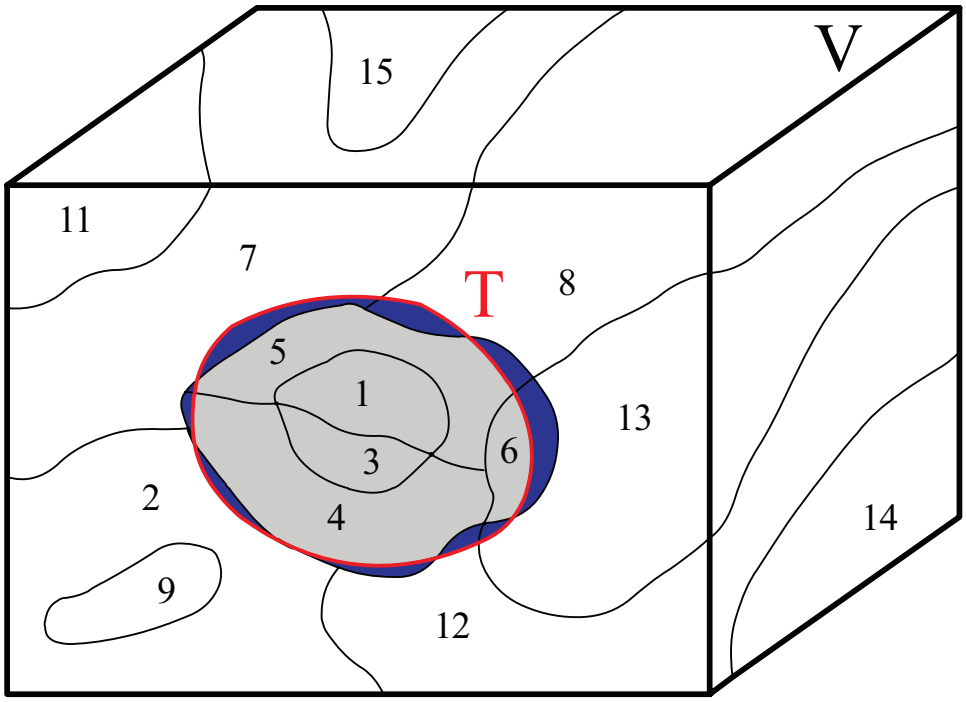
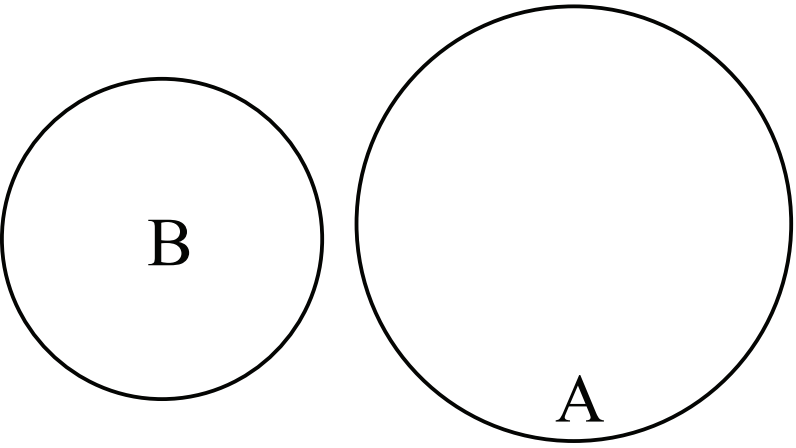


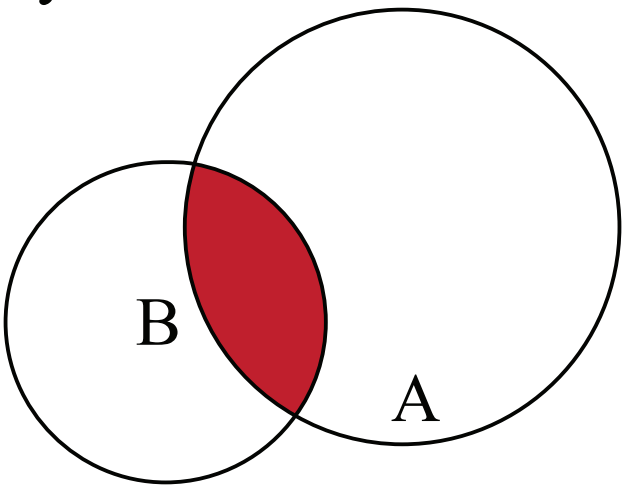
Figure 6



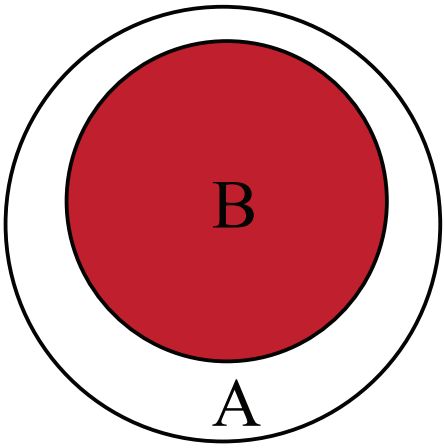
a. RI case study



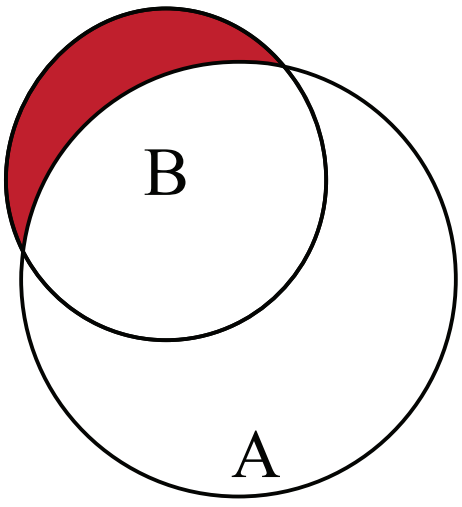
b. $RI(A,B)=0$



c. $RI(A,B)=A\cap B$



d. $RI(A,B)=0$



e. $RI(A,B)=B-A\cap B$

Figure 7

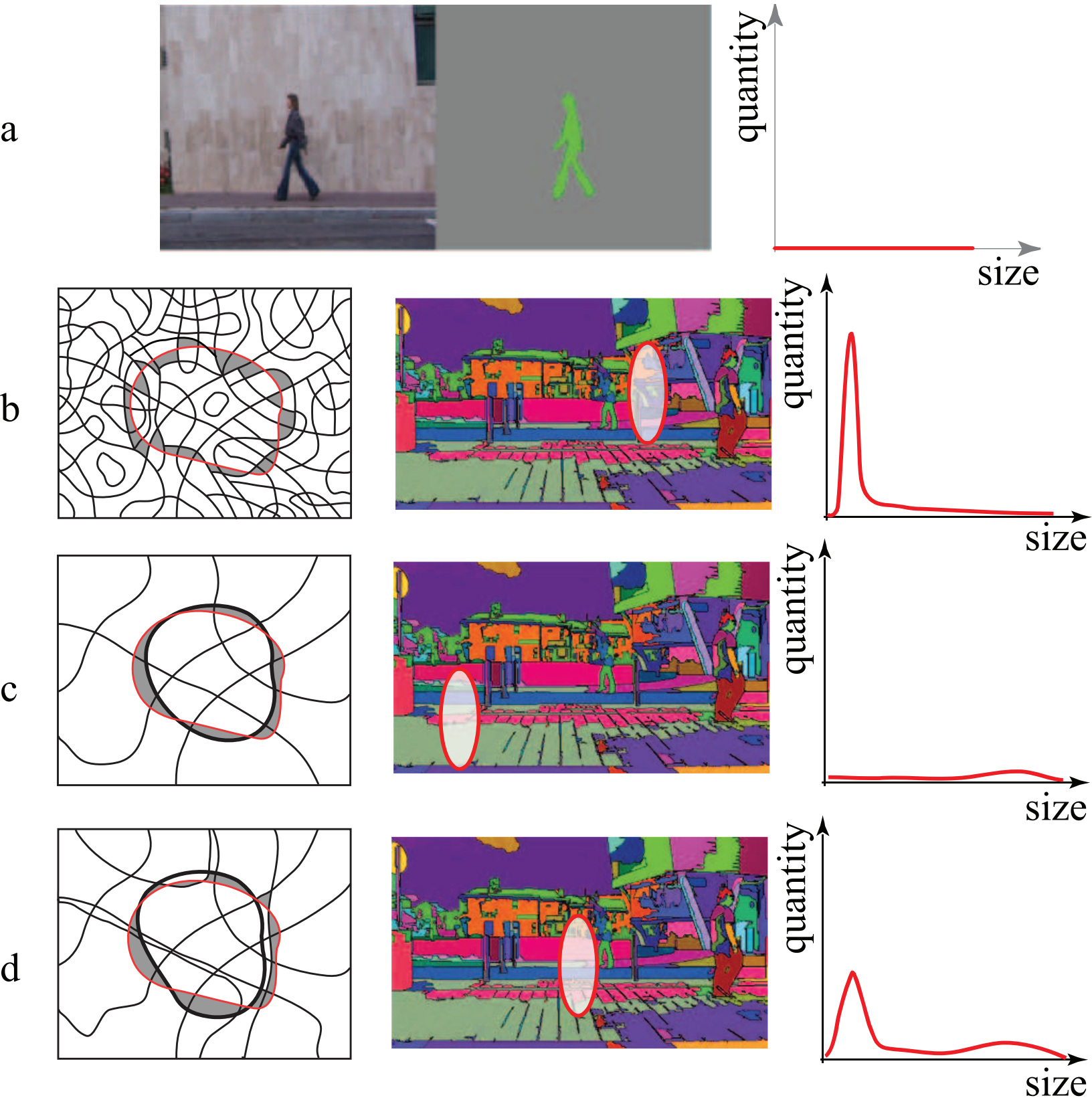
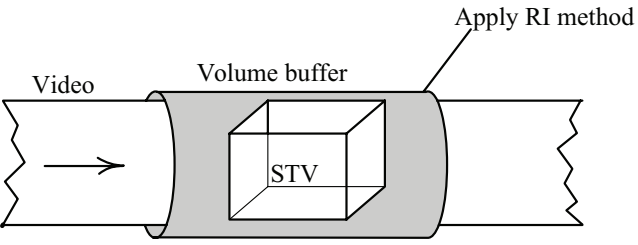
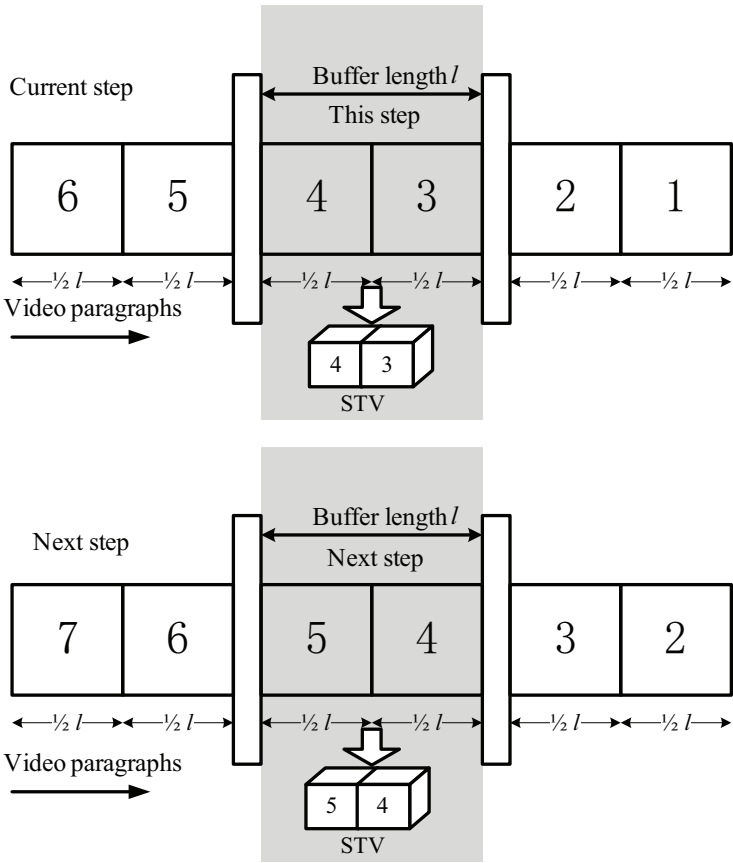

a
quantity
size
b
quantity
size
c
quantity
size
d
quantity
size

Figure 8


Apply RI method
Video
Volume buffer
STV

a


Current step
Buffer length l
This step
6
5
4
3
2
1
½ l
½ l
½ l
½ l
½ l
½ l
Video paragraphs
4
3
STV
Next step
Buffer length l
Next step
7
6
5
4
3
2
½ l
½ l
½ l
½ l
½ l
½ l
Video paragraphs
5
4
STV

b

Figure 9

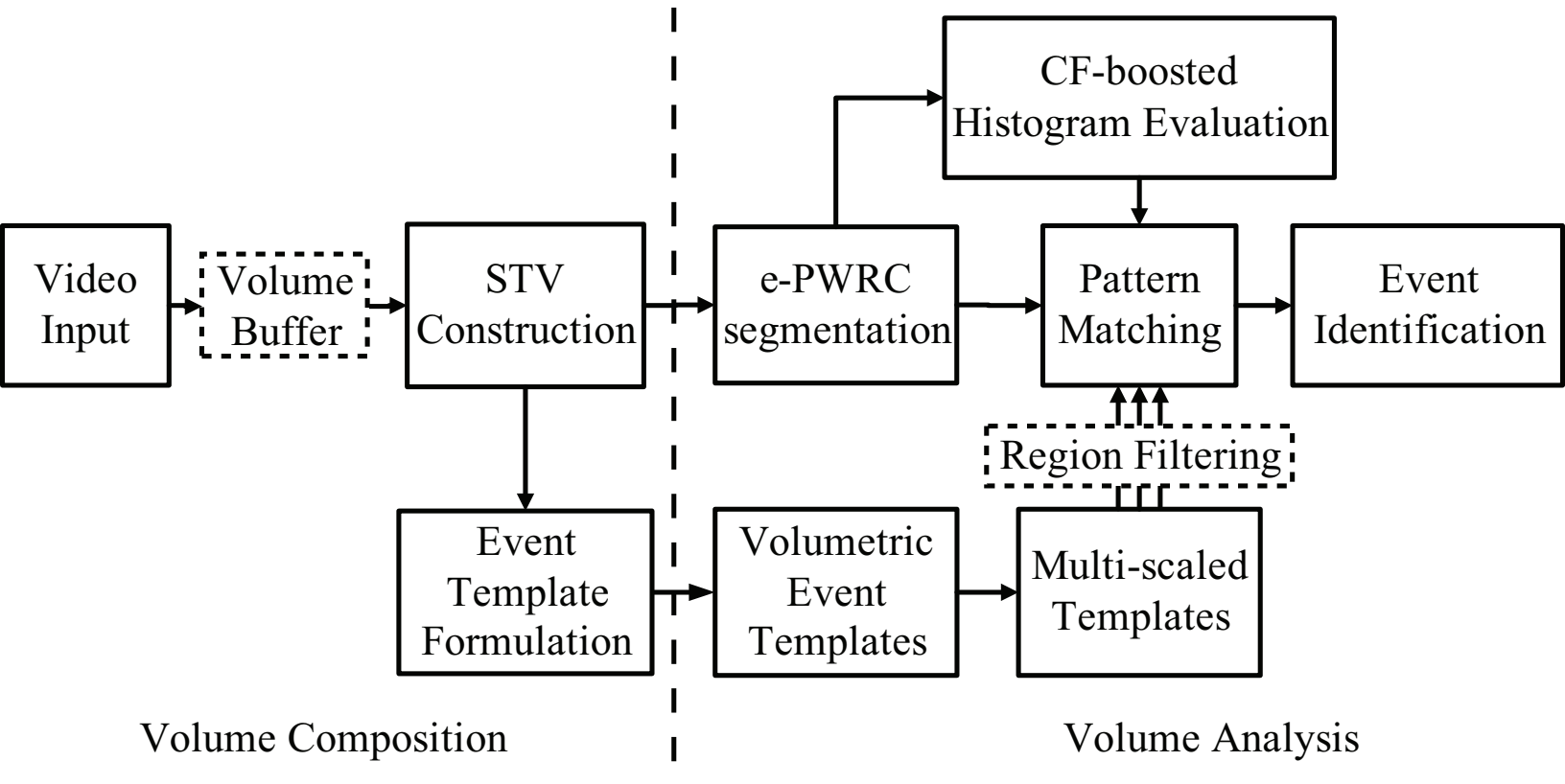

CF-boosted
Histogram Evaluation
Video
Input
Volume
Buffer
STV
Construction
e-PWRC
segmentation
Pattern
Matching
Event
Identification
Region Filtering
Event
Template
Formulation
Volumetric
Event
Templates
Multi-scaled
Templates
Volume Composition
Volume Analysis

Figure 10

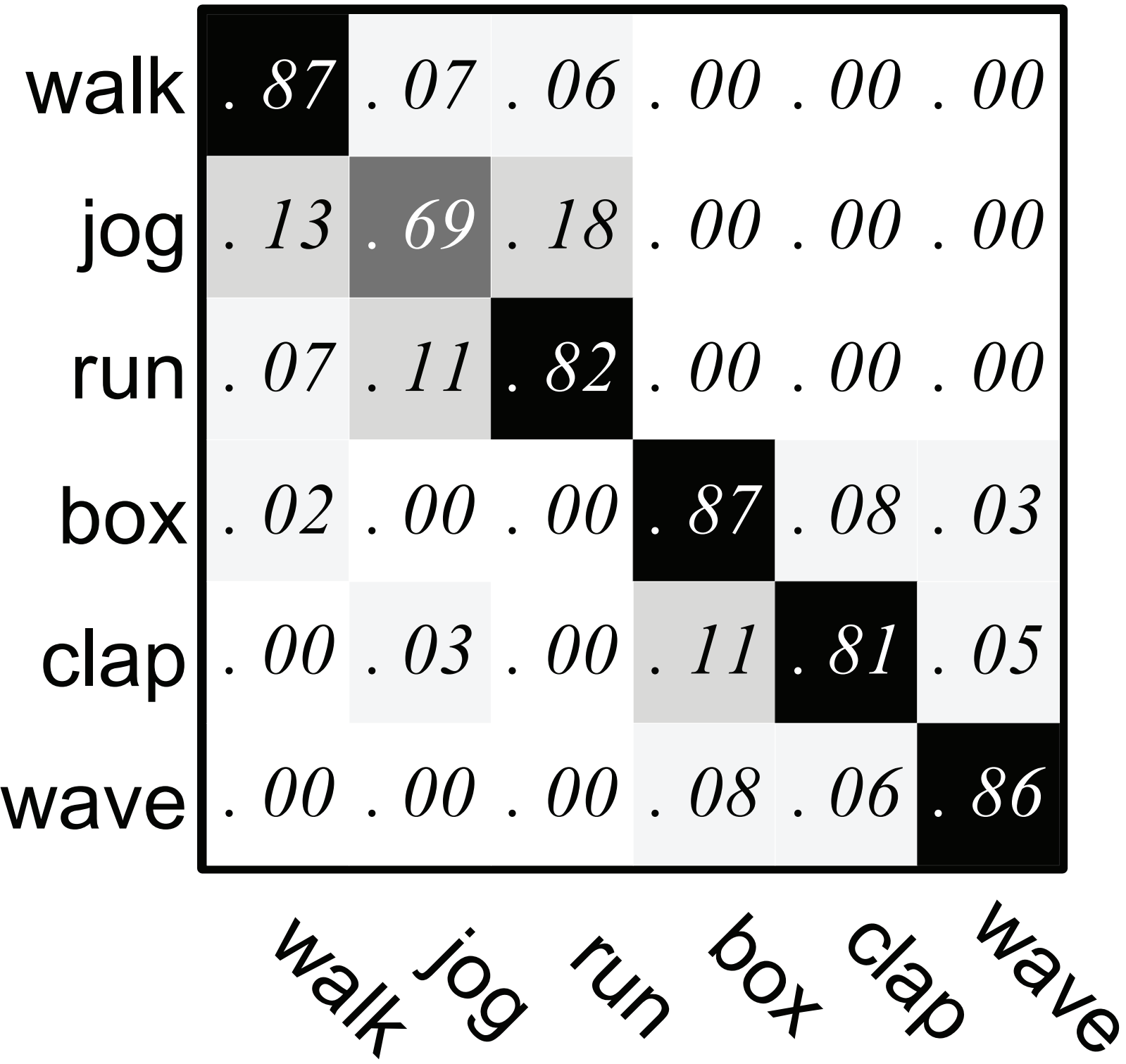
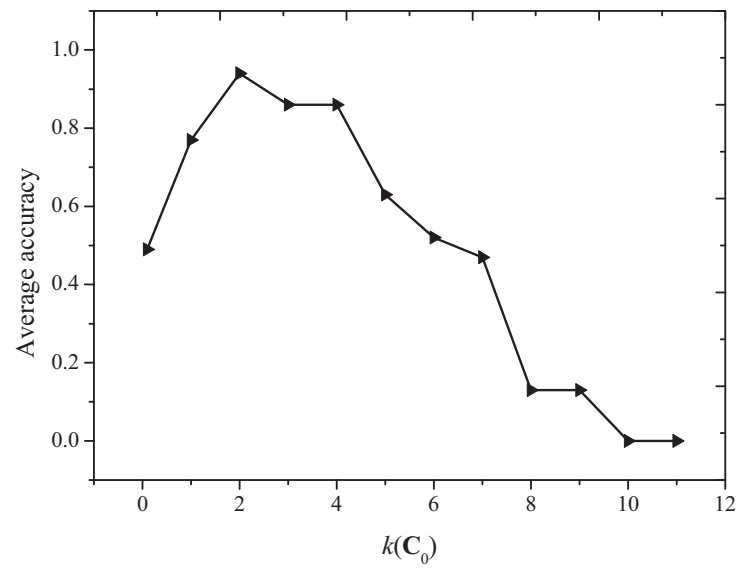
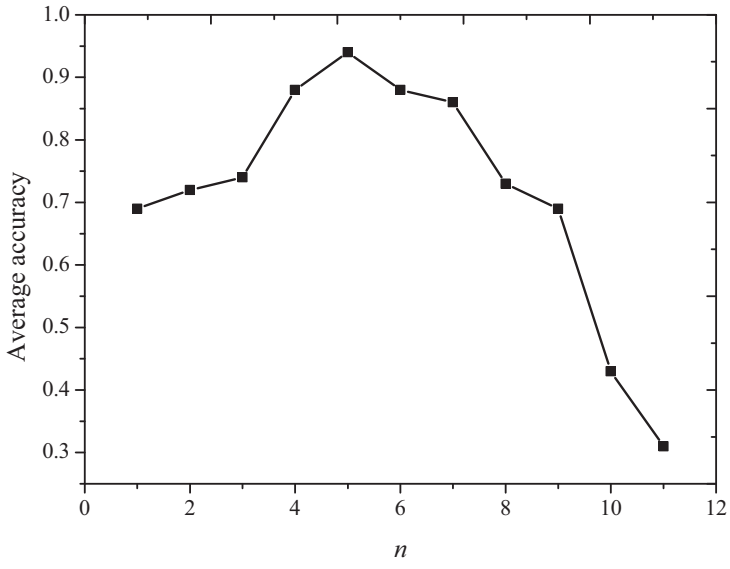
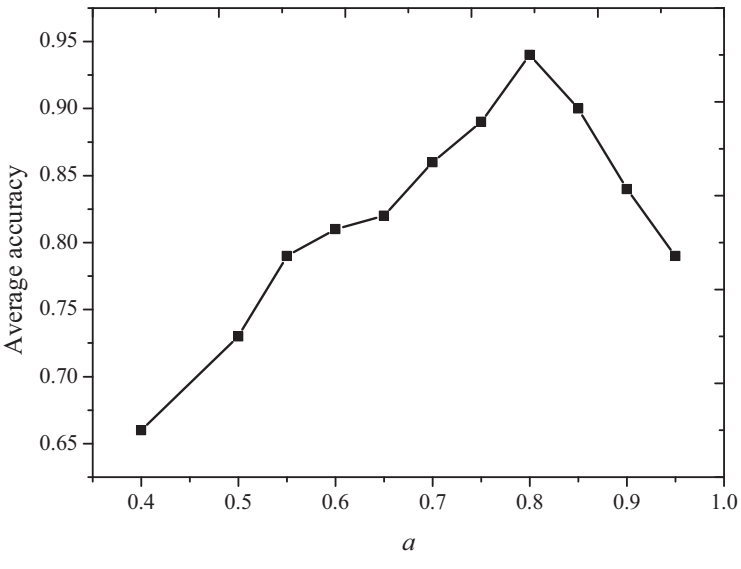

walk
jog
run
box
clap
wave
.87 .07 .06 .00 .00 .00
.13 .69 .18 .00 .00 .00
.07 .11 .82 .00 .00 .00
.02 .00 .00 .87 .08 .03
.00 .03 .00 .11 .81 .05
.00 .00 .00 .08 .06 .86
walk
jog
run
box
clap
wave

Figure 11


Average accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
k(C0)


Average accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
2
4
6
8
10
12
n


Average accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.4
0.5
0.6
0.7
0.8
0.9
1.0
a

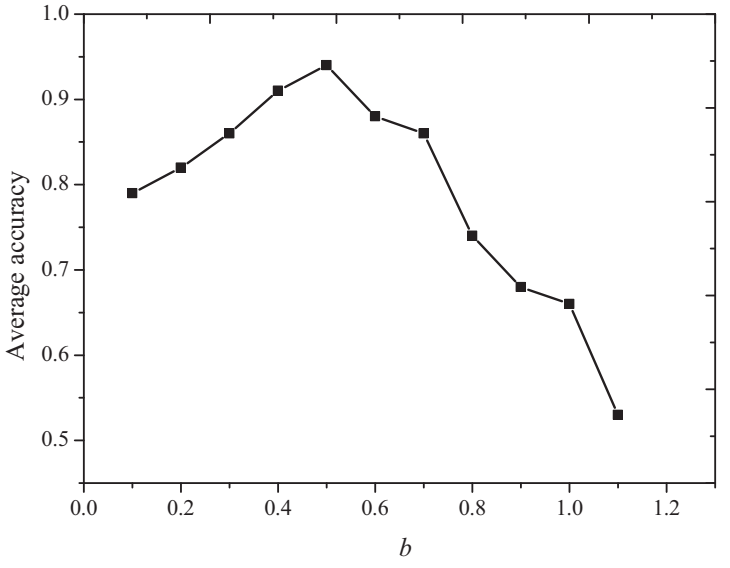

Average accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
b

Figure 12

Click here to download high resolution image

Figure 13

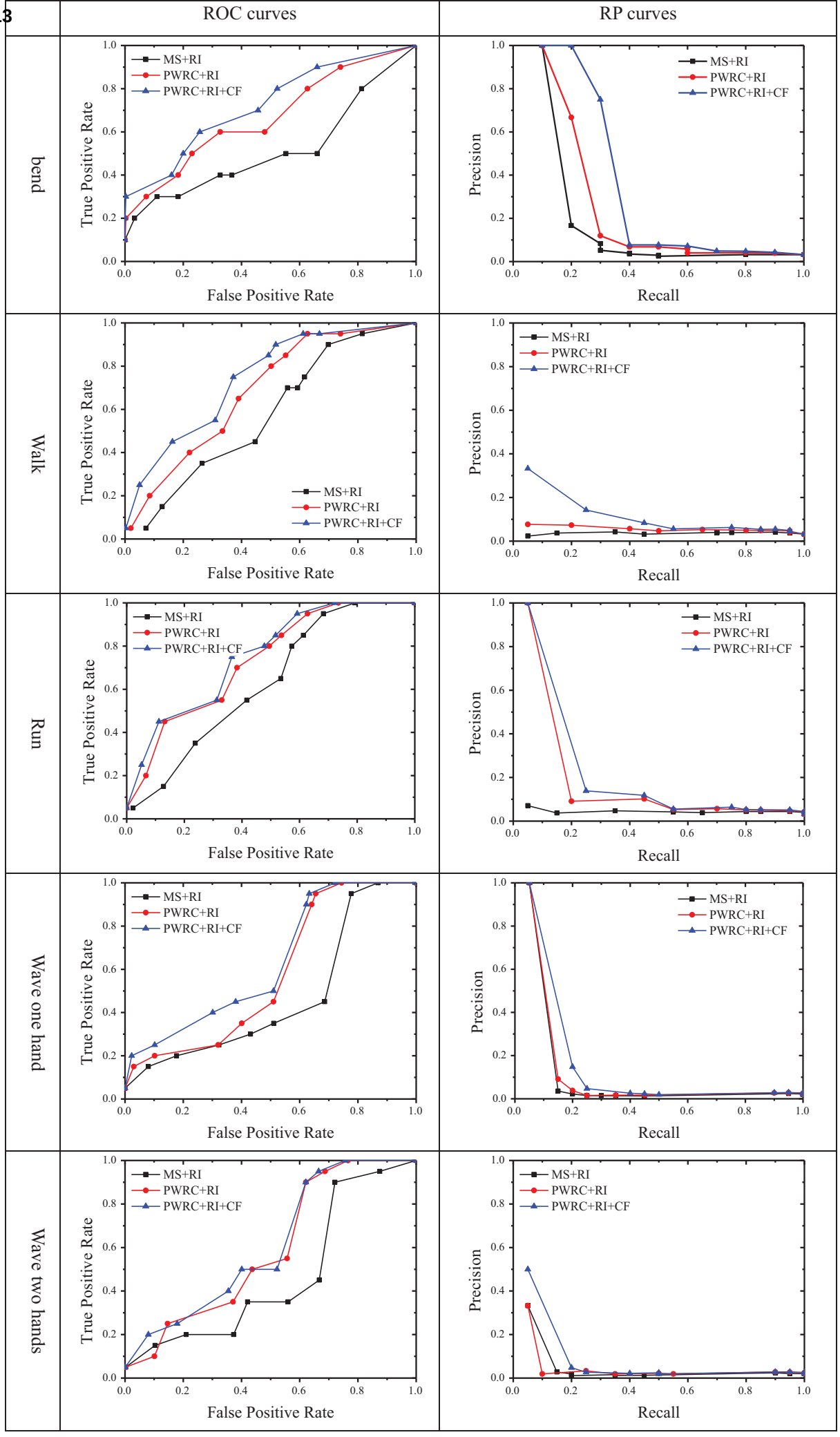
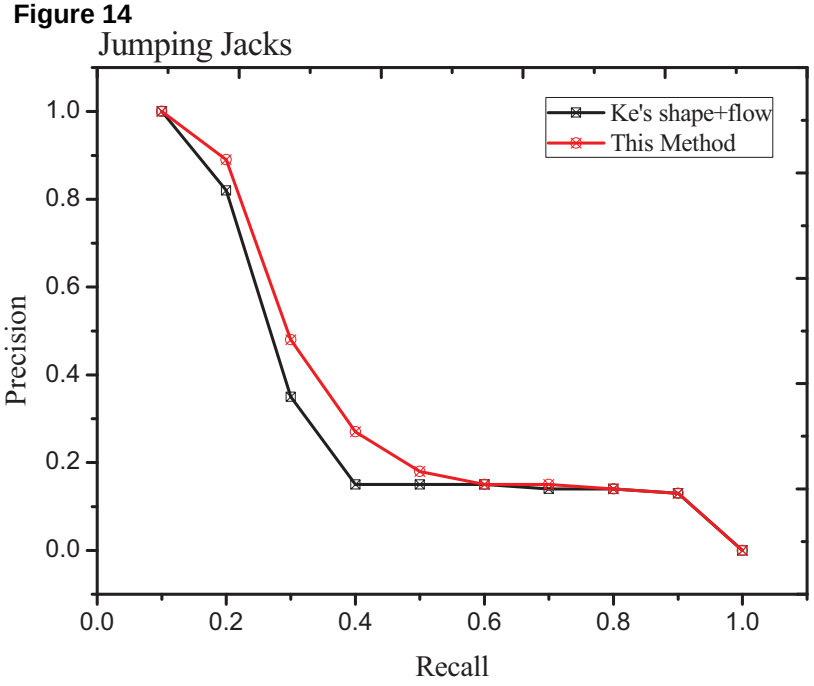
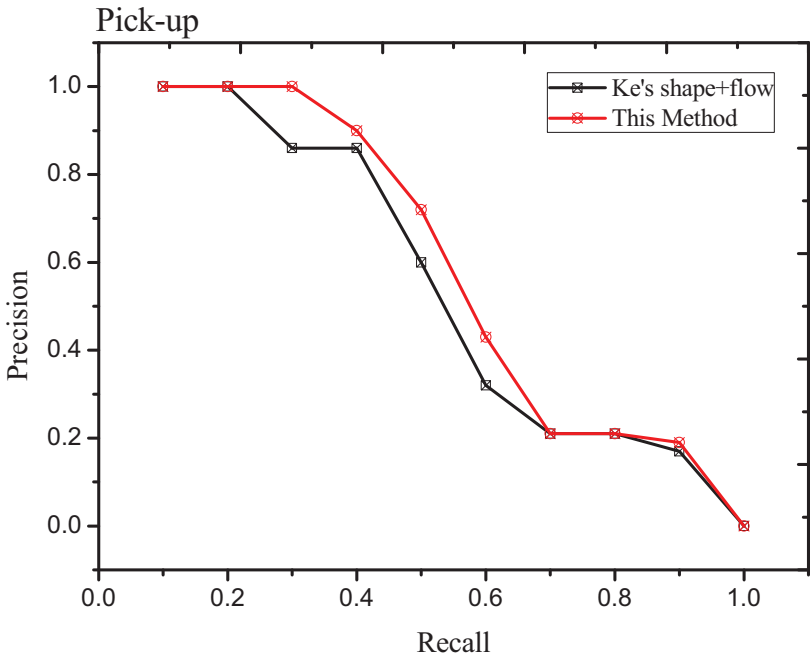
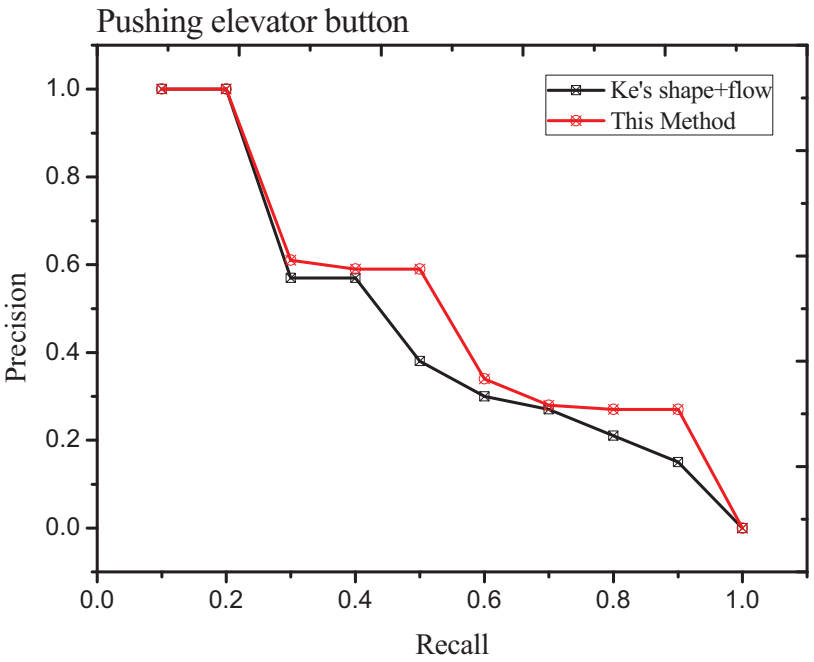
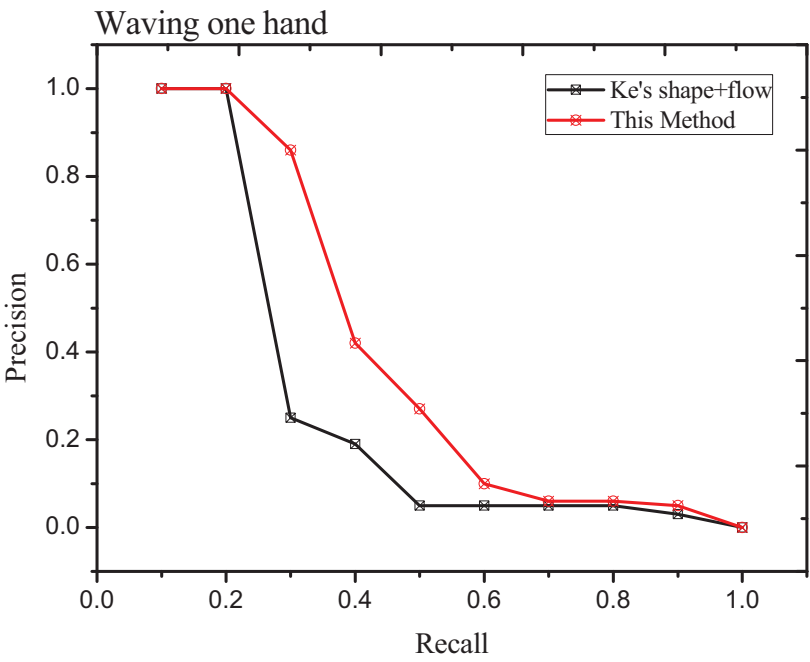

ROC curves
RP curves
bend
Walk
Run
Wave one hand
Wave two hands
True Positive Rate
False Positive Rate
Precision
Recall
MS+RI
PWRC+RI
PWRC+RI+CF

Figure 14


Jumping Jacks
Ke's shape+flow
This Method
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0


Pick-up
Ke's shape+flow
This Method
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0


Pushing elevator button
Ke's shape+flow
This Method
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0


Waving one hand
Ke's shape+flow
This Method
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

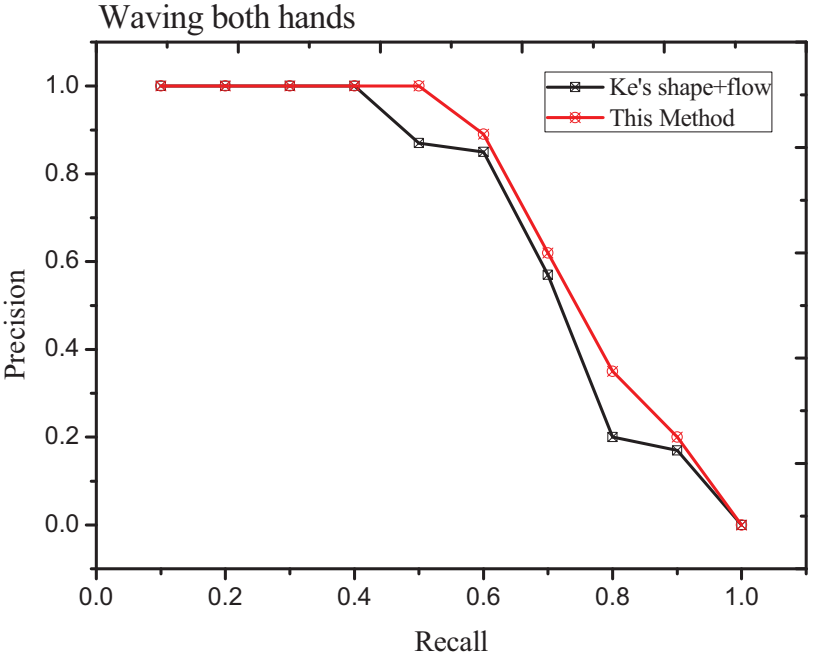
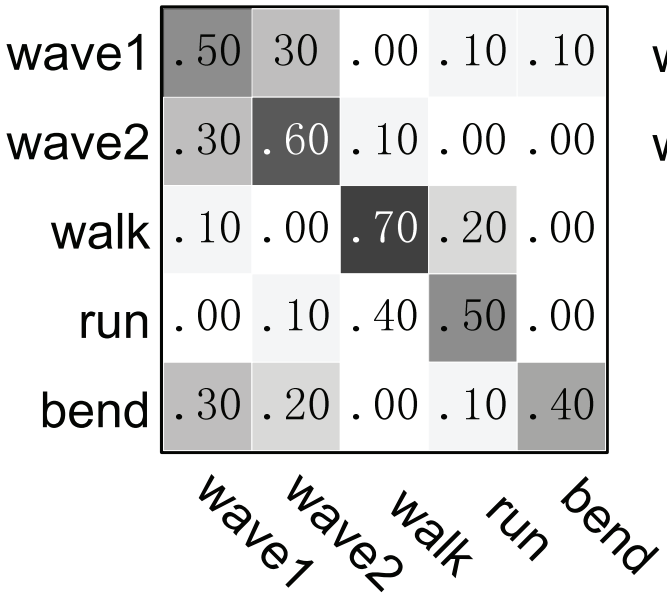

Waving both hands
Ke's shape+flow
This Method
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

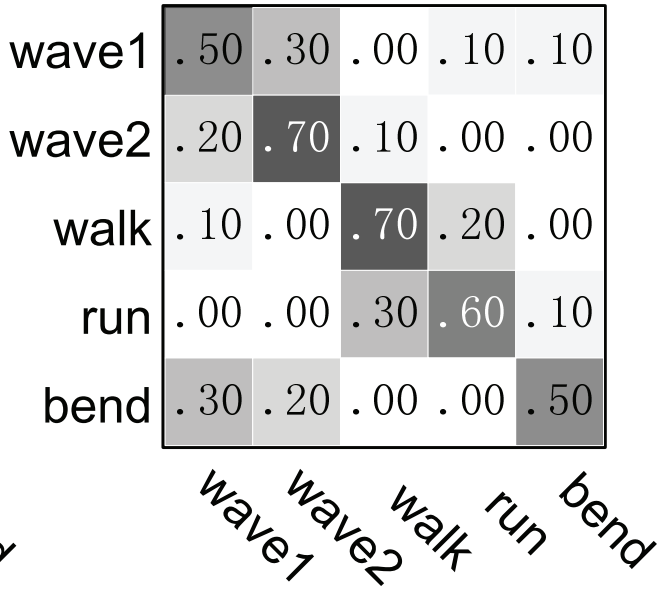
Figure 15

Click here to download high resolution image

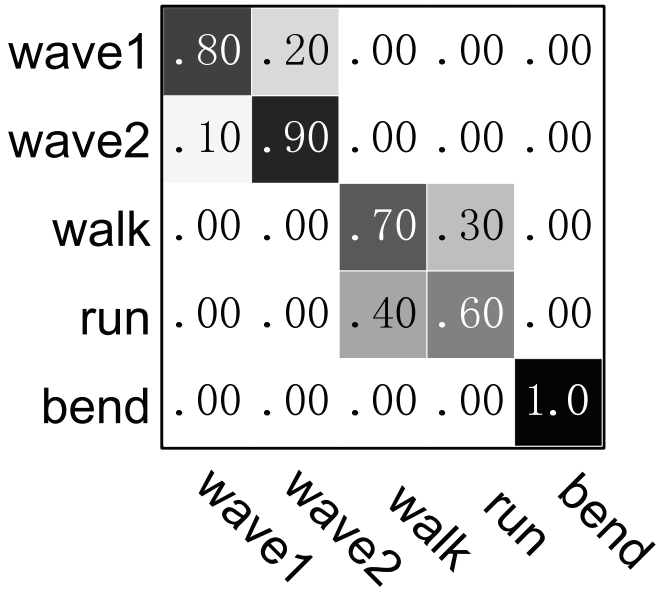
Figure 16



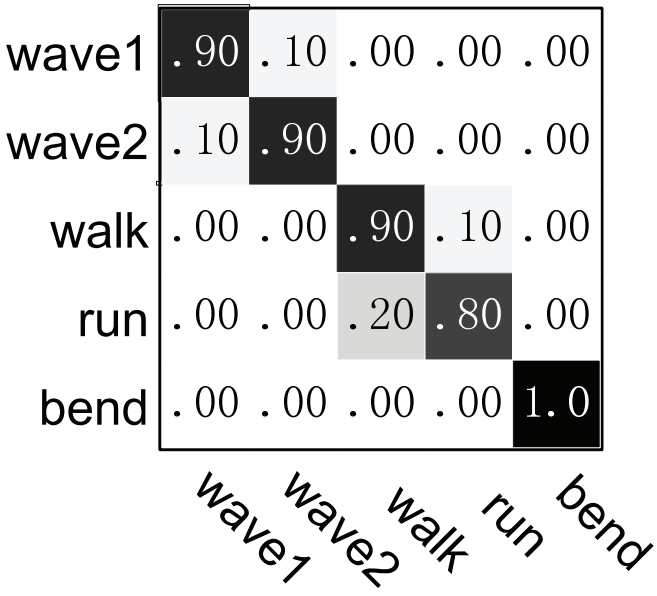
(a) PWRC+RI (without MS)
average accuracy = 57.1%



(b) e-PWRC+RI (with MS)
average accuracy = 61.1%



(c) PWRC+RI+CF (without MS)
average accuracy = 80.2%



(d) e-PWRC+RI+CF (with MS)
average accuracy = 90.1%

Figure 17

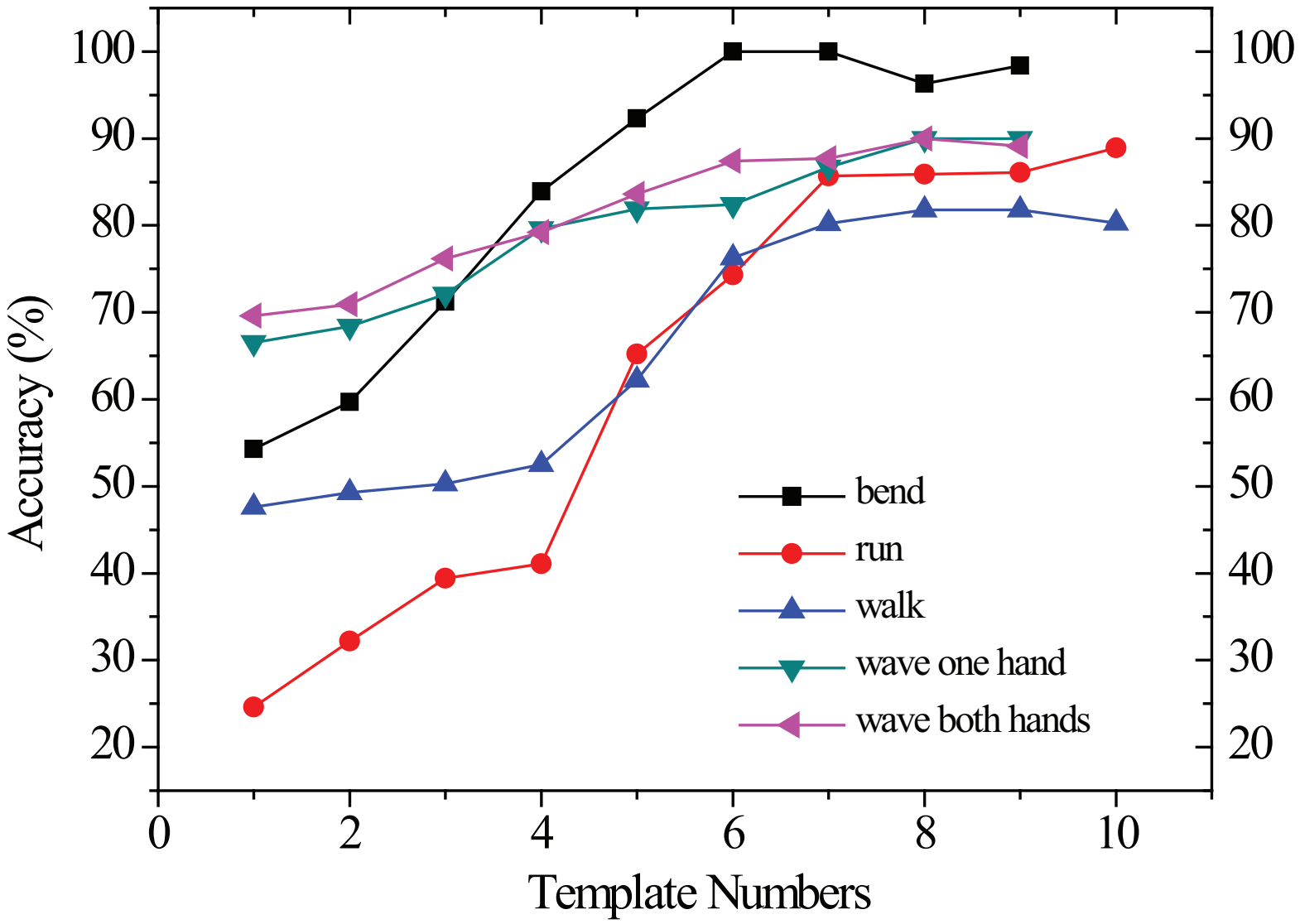

Accuracy (%)
Template Numbers
100
90
80
70
60
50
40
30
20
0
2
4
6
8
10
bend
run
walk
wave one hand
wave both hands

Figure 12

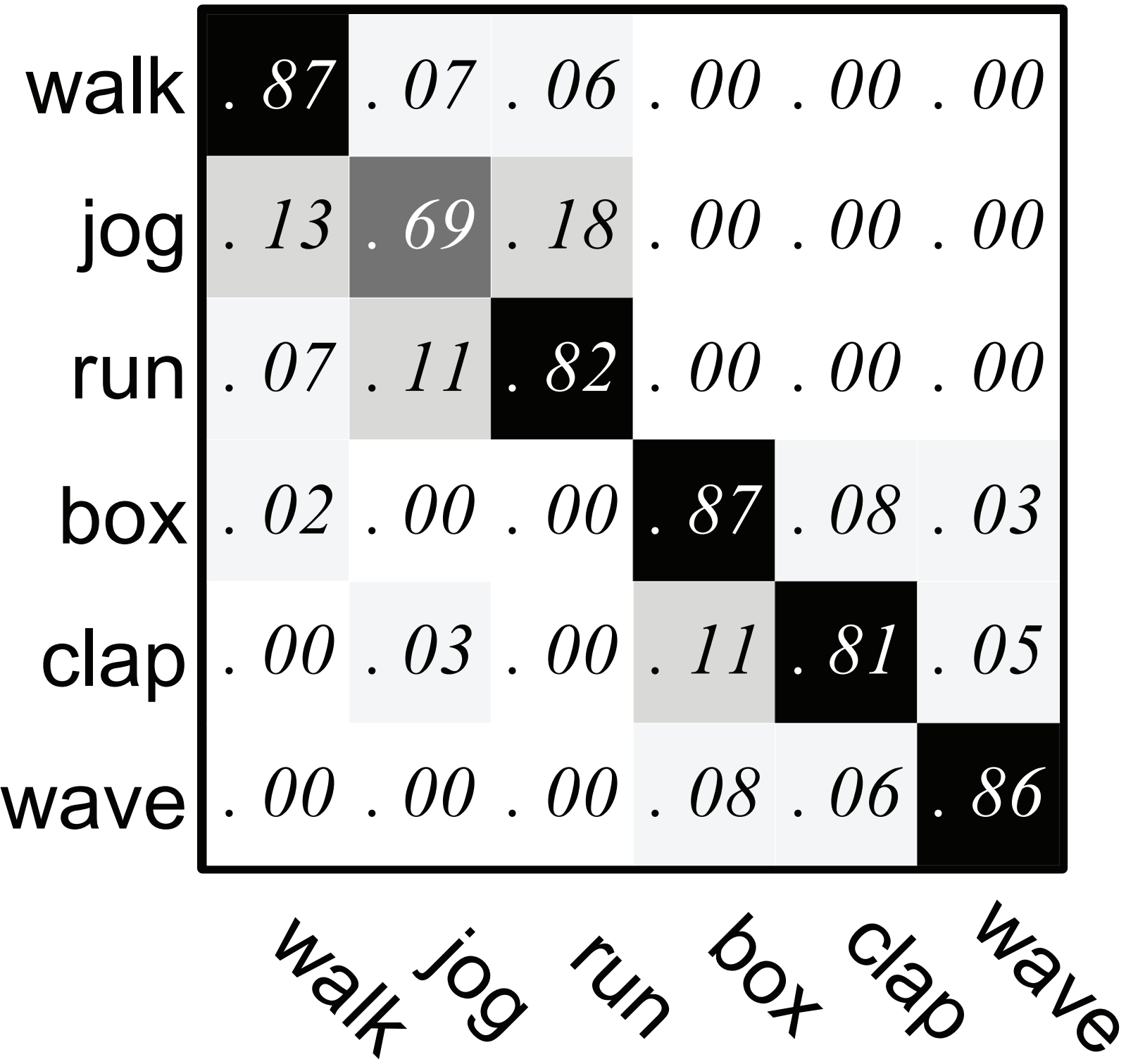

walk
jog
run
box
clap
wave
.87 .07 .06 .00 .00 .00
.13 .69 .18 .00 .00 .00
.07 .11 .82 .00 .00 .00
.02 .00 .00 .87 .08 .03
.00 .03 .00 .11 .81 .05
.00 .00 .00 .08 .06 .86
walk
jog
run
box
clap
wave

Figure 13

Click here to download high resolution image

Figure 14

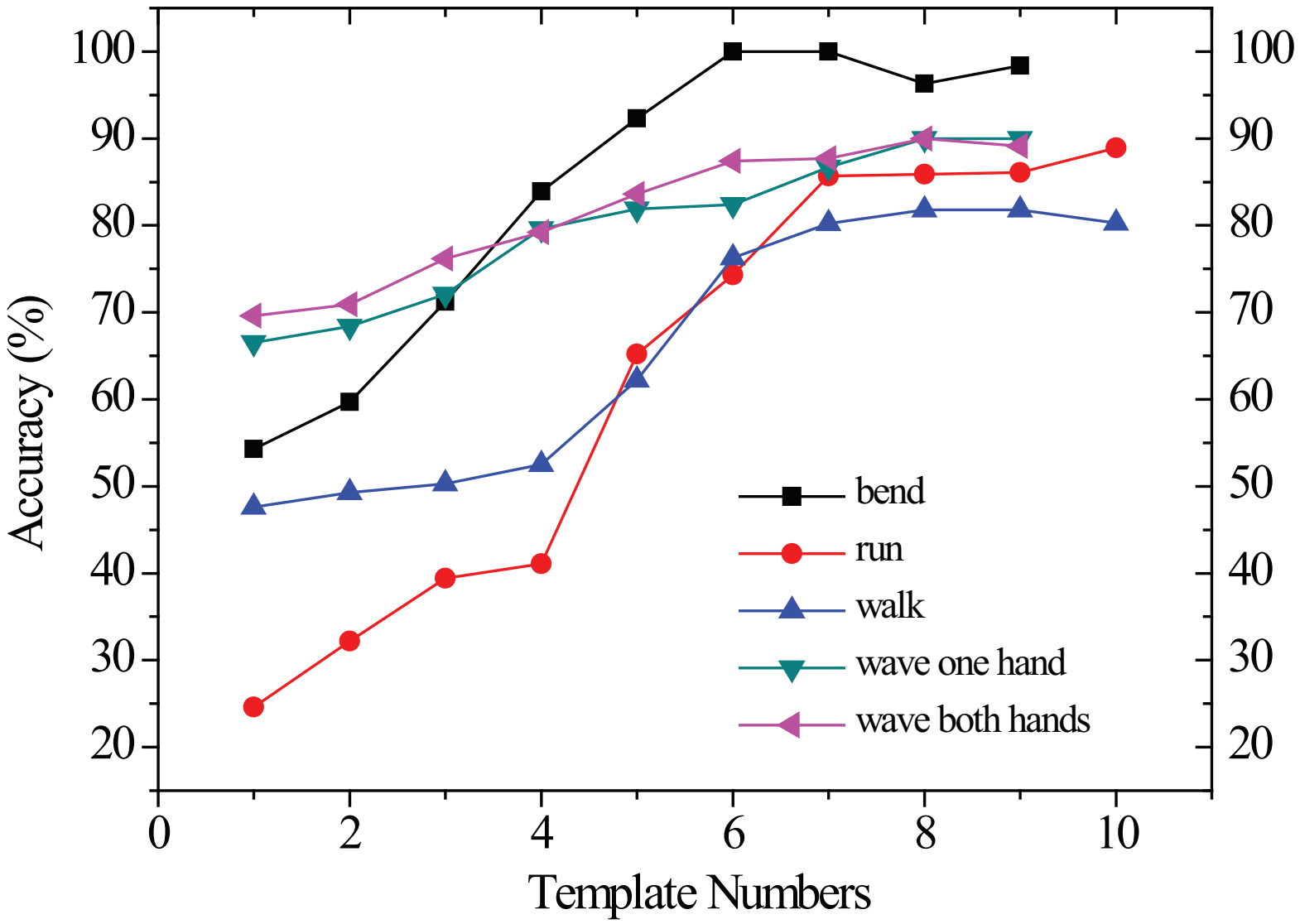

Accuracy (%)
Template Numbers
bend
run
walk
wave one hand
wave both hands
100
90
80
70
60
50
40
30
20
0
2
4
6
8
10

Table 1

Pseudocode Improved Pair-wise Region Comparison

INPUT

1. Spatio-temporal Volume
2. Mean Shift window size factor: h_c and h_l
3. Initialised Pair-wise Region Comparison factor $k(\mathbf{C}_0)$
4. Hierarchical levels n

OUTPUT

STV with Labelled segmentation regions $\mathbf{C}_{n-1}$

ALGORITHM

Initialise $\mathbf{C}_0$ with STV-based Mean Shift segmentation.

Transform STV colour space from RGB to L*a*b*.(Equation 6 to 10)

Loop i from 1 to $n-1$

Build histogram for each region in $\mathbf{C}_{i-1}$

Represent $\mathbf{C}_{i-1}$ as graph:

The vertexes value is L*a*b* colour

The weight of edges is Cha's minimum histogram distances. (Equation 13)

Calculate $k(\mathbf{C}_i)$ (Equation 14)

Calculate $\mathbf{C}_i$ based on original PWRC method (Equation 1 to 4)

Build next hierarchical level on lower resolution (Except the last loop)

End Loop

Output $\mathbf{C}_i$

END Pseudocode

Table 1. Operational Pseudo code for the e-PWRC

Table 2

Pseudocode volumeBuffering (input videoFile)
START:
//Initialization
Allocate appropriate buffer size "***L***" based on videoFile configuration;
Calculate number of time-tablets "***I***" based on event template durations;
Calculate the remainder frames "***N***" at the end of the videoFile;

//Traverse through entire the video volume
Loop (***I***)
    //Calculate the new starting point and length "***R***" of the input STV
        if (at the end of video)
            R = ***N***;
        else
          R = ***L***;

//STV-based template matching
    Release tested STV;
    Compose and renew STV;
    RI matching;
END

Table 2 Volume buffering implementation

Table 3

e-PWRC		Coefficient Factor	
$k(\mathbf{C}_0)$	n	a	b
4	7	0.7	0.7

Table 3 Parameters used for testing KTH dataset

Table 4

Methods and techniques	Event Detection Accuracy
This Method: e-PWRC + RI + CF(Coefficient factor)	**82.0%**
Ke *et al*.'s MS (Mean shift) + RI + Flow [40]	80.9%
Schuldt *et al*. [5]	71.7%
Dollár *et al*. [30]	81.2%
Niebles *et al*. [36]	81.5%

Table 4 The Accuracy performance compared with other approaches

Table 5

e-PWRC		Coefficient Factor	
$k(\mathbf{C}_0)$	N	a	b
5	11	0.6	0.8

Table 5 Parameters applied for uncontrolled testing

Summary of Amendments

Re: Paper No. SIGPRO-D-11-01271

Title: STV-based Video Feature Processing for Action Recognition

Categorization of Reviewers Comments and Recommendations

It is felt that the reviewers had examined the above manuscript carefully and provided the authors in-depth and comprehensive feedbacks. Based on the improvement recommendations, extensive review, expanded referencing, and extra evaluation works have been carried out by the authors. To make the indexing and tracking of the corresponding changes to the manuscript easier, the authors have classified the reviewers' feedbacks into the following categories (Note: R1 stands for Reviewer1 and henceforth R2 and R3):

1. Experiment Design
    a. Expanding the test datasets to evaluate the proposed methods' performances under more complex and challenging conditions; (R2)
    b. Benchmarking the devised techniques against Ke's and other popular action recognition methods; (R1 & R3)
    c. Justifying the appropriateness of the investigated approach for indexing and visual tagging; (R1)
    d. Assessing the independent contribution from the MS-based pre-filtering towards the e-PWRC's overall performance gain; (R2)
    e. Quantifying the parameter sensitivity of the deployed kernel settings. (R2)
2. Paper Structure
    a. Be more explicit on the original contributions from this research; (R1&R2)
    b. Consider a more appropriate title for the paper to accurately reflect the nature of the research; (R1)
    c. Distributing the section weights of the paper more evenly to better focusing on the latest development and research outputs from previous works; (R1 & R2)
    d. Extra literature review on Ke's and Grundmann's methods to justify the improvement of the proposed method; (R2 & R3)
3. References and Writing Style
    a. Citing and analysing more up-to-date and important developments in the field; (R1, R2 & R3)
    b. Reducing non-essential introduction on basic concepts and techniques through more consistent referencing; (R2)
    c. Formal writing and extra proof-reading. (R1)

Revision Specifications

Improved Experimental Design

1. To address Point 1.a and 1.b, the devised techniques have been further tested by applying a wider spectrum of public databases containing arbitrary and more complex backgrounds. Following paragraphs have been added into Section 6:

   For benchmarking the proposed approach, the Recall-and-Precision experiments have been carried out on the widely used video datasets containing complex and dynamic backgrounds [47]. The dataset contains 48 videos and 110 distinguishable human actions labelled as one-hand waving, two-hand waving, picking, pushing, and jumping-jacks, etc. Figure 14 shows the detecting accuracies denoted by relevant RP curves. by, The proposed method and its corresponding techniques have shown more consistent and improved performance in comparisons with the parts-based template matching and flow correlation method [40].

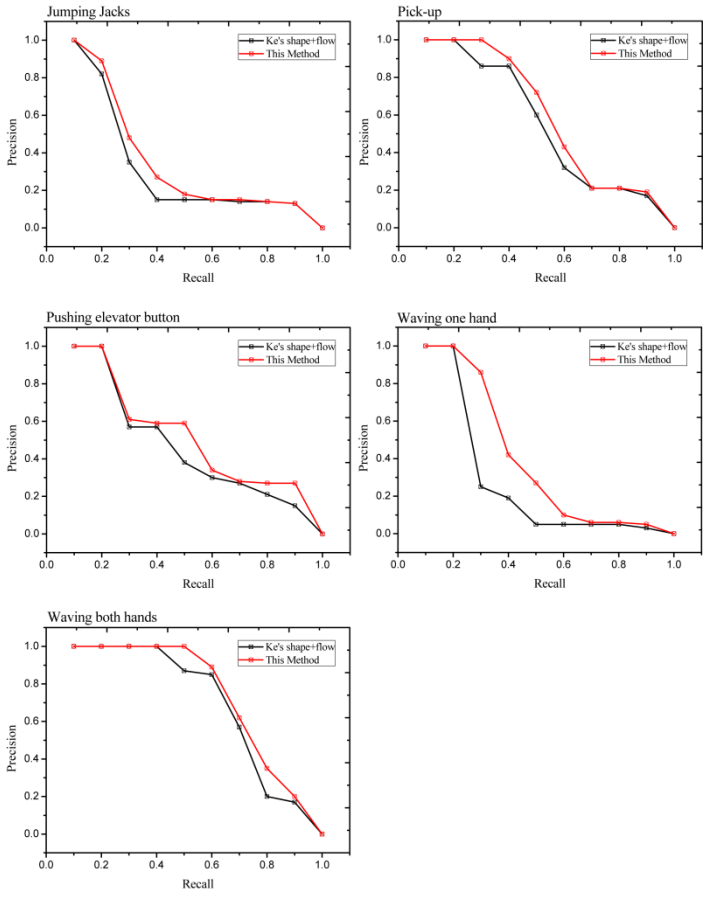


Figure 14. Benchmarking Recall-and-Precision experiments on complex video datasets

2. To address Point 1.c, real-world CCTV recordings have been tested for video indexing and visual tagging applications as detailed in Section 6.2:

   For testing the prototype system's performance on visual tagging and video indexing, a set of real-world CCTV files have been utilized. The original video clips were downloaded from the social website – YouTube [48]. Figure 15 shows snapshots and detection result from the videos showing a number of pedestrians tripping and falling down at a spot near the entrance of a building. The developed system has been used to detect and denote all the "tripping-and-falling" events from the footages. The average length of the input videos is approximately 5 minutes with 8 "tripping-and-falling" events. The first 4 events are used for defining the template. All the remaining 4 events were successfully identified and denoted. It is clearly visible from the snapshots that the original videos were filled with noise signals and containing moving objects such as vehicles and pedestrians in the scenes.

Figure 15. Detecting and indexing exercise of the "Falling down" events. The templates are highlighted in the first row. Selected snapshots from video feed-ins are lined up in the second row, where the ground truths are denoted by the red overlaying areas, and the indexed results highlighted by the orange regions.

3. The overall performance of the developed system has been benchmarked against other published counterparts to reflect on the Point 1.b and 2.c. The results are detailed in Section 6. Additional analyses on the distinctive characteristics of the devised approach have been added at the end of Section 6.1. For example, for justifying the performance variations of the proposed method when applied to a particular dataset – KTH, following paragraph has been added:

Compared with some recent published methods, such as [37, 46], whose accuracies registered on the KTH datasets often reached near 90%, the proposed method seems produced a slightly lower detection rate. This is because the proposed approach had been focusing on more generic application scenarios for recognizing everyday human activities without strong assumptions and pre-conditions on the settings of backgrounds scenes and human models, which offers improved consistency and robustness for real-world implementation.

4. To address Point 1.d, extra experiments have been conducted (see amended Section 6.3). The independent contributions of the Mean Shift pre-clustering process have been evaluated. The new series of experiments have also served as an integral part of the improved assessment strategy of this programme.

   It is evident in the experiments that the e-PWRC method has been benefitted significantly from its MS-based hierarchical design and the dynamic parameter controls. Similar conclusions were reached by comparing the e-PWRC-empowered method with the Coefficient Factor (CF)-boosted RI method as shown in Figure 16. The event detection accuracy has registered a 9% increase by using the devised e-PWRC framework.

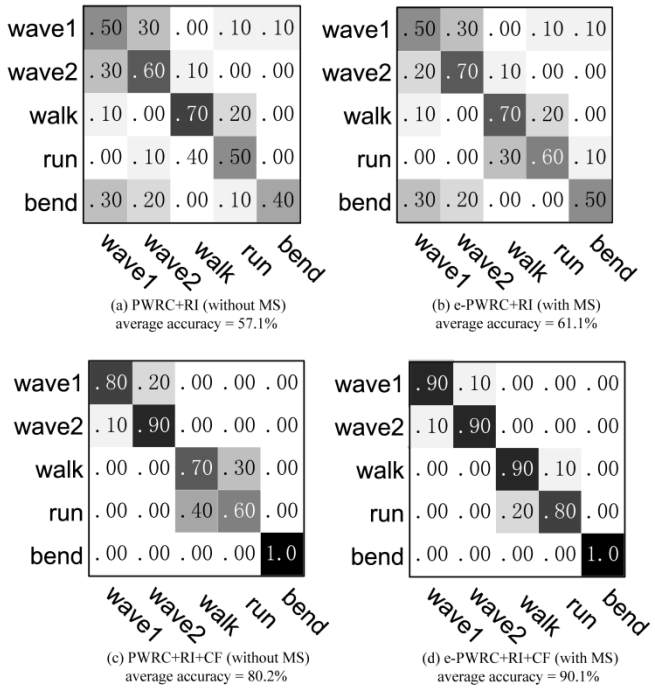


Figure 16. The confusion matrices on the campus dataset for evaluating the independent MS contribution within e-PWRC.

Different from Grundmann's approach [42], the vertices of the initial PWRC graph and k($\mathbf{C}_0$) in the Equation 14 in the new method were based on the aforementioned MS clustering results rather than the direct employment of the original colours of each voxel. The improved k($\mathbf{C}_0$) corresponds to the MS parameters, H_c and H_l, to highlight the importance of the low-level colour features during the initial PWRC merging calculations at the bottom layer within the hierarchical structure. After the first merging operation is completed, the PWRC segmentation will literally shift from a colour-based process to texture-based calculations. The experiment results have shown that during the feature extraction stage, the MS pre-clustering actually serves as an effective measure for simplifying the initial graph setting.

As indicated by the experimental results, the e-PWRC method has proven its effectiveness for providing quality outputs for the following RI matching operations, especially when subjecting to complex real-world conditions. Accompanied by the optimisation measures employed at the template matching stages, the process pipeline and its various operational models have shown satisfactory consistency and robustness.

5. To address Point 1.e, a new experiment has been designed for testing the sensitivity of each parameter within the CF-boosted RI distance model. Following conclusions have been added in Section 6.1.

   An experiment has been carried out for evaluating the sensitivity of the four parameters utilised in the devised system approach, the e-PWRC factor k($\mathbf{C}_0$), the hierarchical level n, and the linear factors of the RI coefficients a and b. The same templates employed in the previous accuracy evaluations from each action event categories have been utilised. The independent contributions from each of the 4 parameters have been evaluated by fixing the other three at their optimized value. Figure 11 denotes the relationships between each parameter and the average detection accuracy.

   It is noticeable in Figure 11, the curve patterns of the e-PWRC-related parameters k($\mathbf{C}_0$) and n have shown similar distributions while the e-PWRC transmitting from over- to under-segmentation. Hence, the system performance

can be improved significantly if more accurate over-segmented sub-regions can be extracted. It is also worth noting that the purposed method can still successfully detect about 42% of the queried events even under extremely over-segmented conditions.

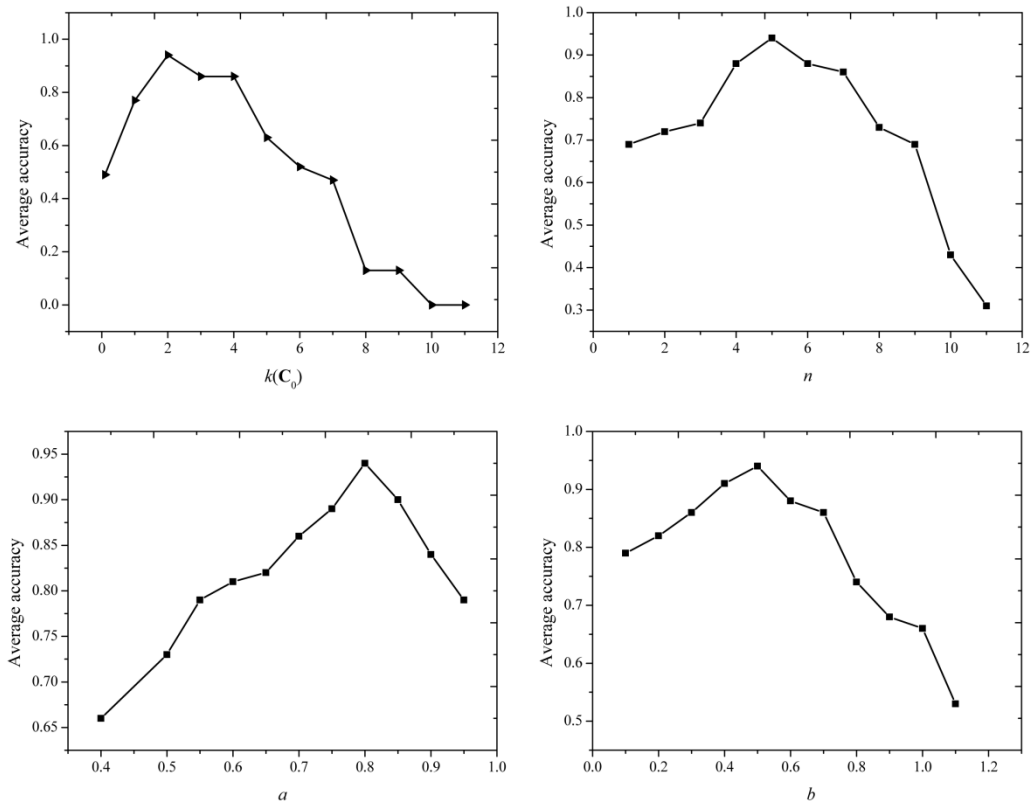


Figure 11. Parameter sensitivity test result

As discussed in Section 4.3, the values of the coefficient factor a and b should be around 1 based on the normalized χ^2 distance of the histograms, which has been set as the threshold in this design for switching between "rewarding" and "punishing". In Figure 11, the accuracy performance is above 90% for detecting the "waving" when appropriate a and b have been applied to the threshold.

Revamped Paper Structure

For providing more explicit rationales and highlighting the original contributions from this research, the revised paper has introduced a number of new sections and paragraphs, as well as removing substantial quantity of less significant background information.

1. The main contributions from this research have now been specified and clearly stated at the start of the paper, as suggested by reviewers (see Point 2.a):

   This paper will focus on two main contributions from the research to the action recognition domain:

   - *An innovative segmentation algorithm for extracting voxel-level features has been developed based on a hybrid discontinuity and similarity-based segmentation model through combining Mean Shift (MS) clustering and the graph-based region description method.*
   - *A novel extension of the Region Intersection technique has been realized for human action recognition by using a coefficient factor-boosted template matching method; that is capable of superior performances under complex and real-world videoing conditions.*

2. The phrase "event detection" that was extensively used in the original manuscript encompasses an extensive spectrum of computer vision topics. For better alignment with the intrinsic characteristics of this research and paper, it has now been replaced by a more specific phrase - action recognition. For example, in Section 1, the following paragraph has been added:

   In this research, a spatio-temporal volume (STV) and region intersection (RI) based 3D shape-matching method has been proposed to facilitate the definition and recognition of human action events recorded in videos.

   Also, the title has now been changed to "*STV-based Video Feature Processing for Action Recognition*" to respond to the reviewer's recommendation in Point 2.b.

3. To address Point 2.c, the abstract has been re-written and the conclusions shortened:

   Video recordings can provide rich and intuitive information on dynamic events occurred over a period of time such as human actions, crowd behaviours, and other subject pattern changes in comparison to still image-based processes. However, although substantial progresses have been made in the last decade on 2D image processing and its applications such as face matching and object

recognition, video-based event detection still remains one of the most difficult challenges in computer vision research due to the wide range of continuous or discrete input signal formats and their often ambiguous analytical features. In this paper, a spatio-temporal volume (STV) and region intersection (RI) based 3D shape-matching method has been proposed to facilitate the definition and recognition of human actions recorded in videos. The distinctive characteristics and the performance gain of the devised approach stemmed from a coefficient factor-boosted 3D region intersection and matching mechanism developed in the programme. This research has also investigated techniques for efficient STV data filtering to reduce the amount of voxels (volumetric-pixels) that need to be processed in each operational cycle in the implemented system prototype. Encouraging features and improvements on operational performance have been registered in the corresponding experiments.

4. To address Point 2.d, the improvement of the proposed action recognition approach from the baseline method introduced by Ke *et al*. has been specified in Section 2.3:

   Based on the spatio-temporal shape feature distribution and the local region grouping, an RI-based distance algorithm harnessing the advantages of both global and local features for recognizing human actions was first introduced by Ke [40] in 2007. Although proven effective when segmenting backgrounds of average complexity, the algorithm was less useful when dealing with complex real-world scenes where both extremely large and small textured regions existed alongside each other. This was one of the early motives of this research to investigate system approaches with more consistent performance for real-world applications.

Additional References and Proofreading

1. To address Point 3.a, additional and up-to-date publications in the field have been referenced and discussed as recommended by the reviewers. For example, in Section 2.3, following paragraphs have been added:

   As summarized by Shao [27, 28], a recent trend of integrating and enhancing STV global features with a chosen set of local features for calibration has

witnessed a degree of success. For example, Jiang et al. [29] had introduced an inter-frame constrained local feature definition method for creating a so-called "convex matching scheme" to facilitate human action detection. Dollár et al.'s [30] spatio-temporal cuboid prototyping method had extended the ability of the 2D interest point segmentation technique into Basharat et al.'s [31] SIFT-based (Scale Invariant Feature Transform) video retrieval pipeline; Zhai and Shah's [32] spatio-temporal attention model, Laptev and Lindeberg's [33] slice-based feature identification techniques, Loper et al.'s [34] bag-of-visual-features (BoVF) methods, as well as Shao's motion and shape feature-based video content segmentation strategies [35] have all been pushing boundaries on this front.

Recently, action recognition researches have been experiencing a weight-shift from the low lever feature and algorithm-oriented development to the semantic-informed machine learning domain. For example, the local feature based Bag of Words (BoW) [36] algorithm and its variations [37] have been used in machine learning-based classification through the sliding window mechanism. Similar efforts, such as Hu's oriented gradients feature with support vector machines (SVMs) [38] technique and Yang's motion video patch distance with cascade classification method [39] had partially addressed the occlusion problems often occurred in real-world video recordings

2. Also addressing Point 3.a, the Grundmann's 2010 paper has been cited with a discussion added in Section 3.1:

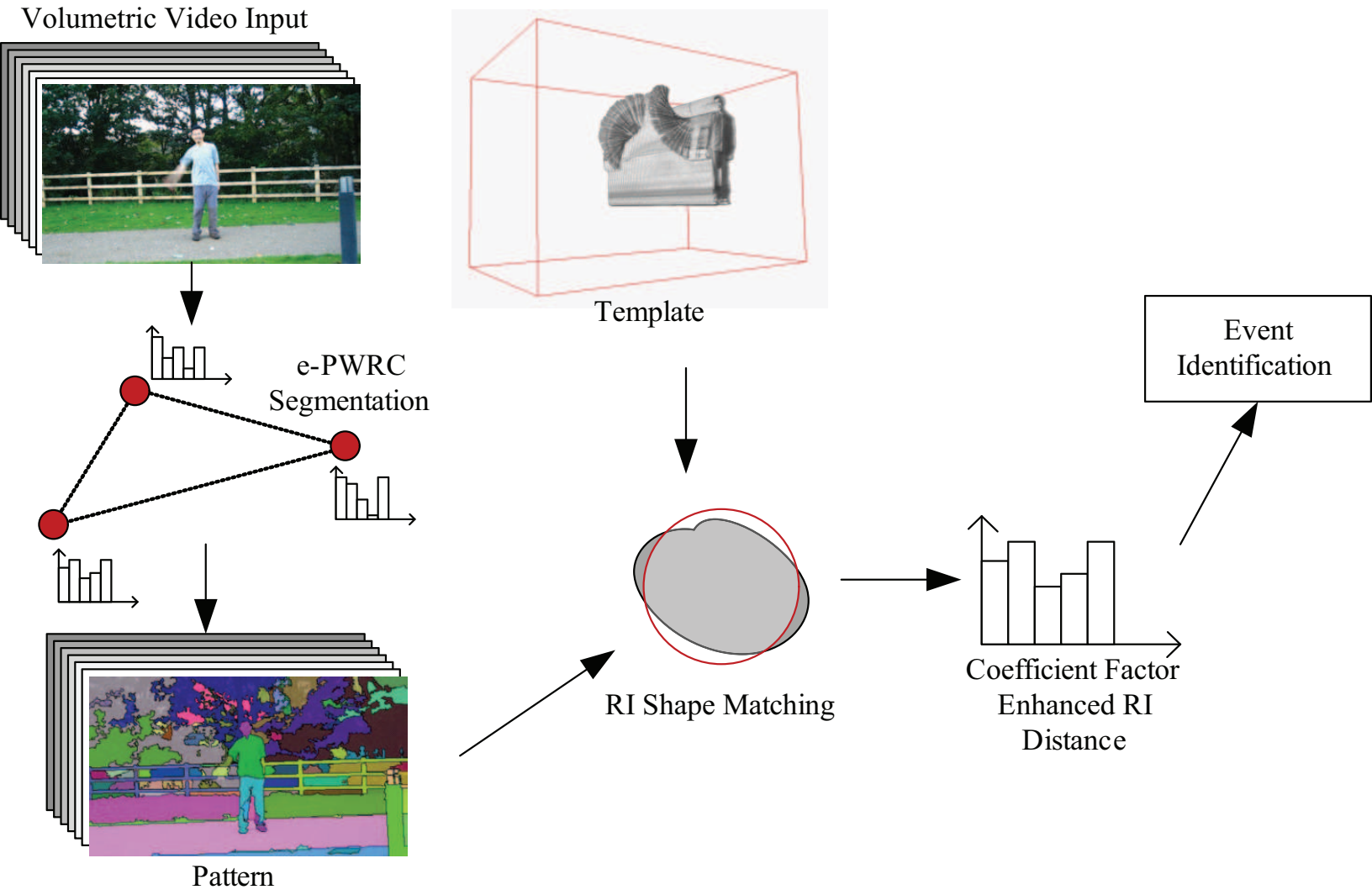
   The proposed solutions for the aforementioned problems were partially motivated by Grundmann's earlier work on video segmentation [42], which focused on the video region description by using texture similarity. It highlighted the importance of similar regions along the temporal domain. To further strengthening the local voxel colour characteristics of the proposed graph-based algorithm, this research has introduced a Mean Shift-based pre-clustering operation for constructing the initial region groups that emphasized the colour features in addition to Grundmann's texture features at the beginning of the e-PWRC operations.

3. To address Point 3.b, non-essential but related concepts and background information have been removed from the main text and referred in appropriate

manner. For example, details of the 2D PWRC method introduced in the original manuscript (Section 3.1) has become a brief review over the parameter *k* used in the e-PWRC algorithm. More details can be referred to the authors' earlier paper. The entire section about the Lab colour space has been removed from the paper. The paragraphs discussing the non-human based video events have been removed from Section 2.1.

4. The entire manuscript has gone through additional rounds of proofreading to respond to Point 3.c.

Graphical Abstract



Highlights

- Video event definition based on volumetric and over-segmented shape regions
- Advanced 3D feature extraction approach based on improved Pair Wise Region Comparison
- Innovative event template matching using Region Intersection techniques
- Optimized implementation strategy through an on-fly buffering mechanism