The 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real-World Applications
In Conjunction with 13th International Conference on Computer Vision (ICCV 2011)
Oral Presentation (Camera Ready)

Video Event Detection Based on Over-segmented STV Regions

Jing Wang, Zhijie Xu
Computer Graphics, Imaging and Vision Research Group
School of Computing and Engineering
University of Huddersfield, Huddersfield HD1 3DH, United Kingdom
{j.wang2, z.xu}@hud.ac.uk

Abstract

Real-world environment introduces many variations into video recordings such as changing illumination and object dynamics. In this paper, a technique for abstracting useful spatio-temporal features from graph-based segmentation operations has been proposed. A spatio-temporal volume (STV)-based shape matching algorithm is then devised by using the intersection theory to facilitate the definition and detection of video events. To maintain system efficiency, this research has integrated an innovative feature-weight evaluation mechanism which "rewards" or "punishes" recognition outputs based on the segmentation quality. Substantial improvements on both the event "Precision" and "Recall" rate and the processing efficiency have been observed in the experiments in the project.

1. Introduction

Video event detection is an application for finding and interpreting real-world activities through appropriate image feature extraction and recognition processes. As one of the hotly-pursued computer vision research areas, video event detection and its relevant techniques have been used for understanding human actions, crowd behaviours, and other non-human-oriented patterns based on real-time or off-line video footages. A number of successful pilots on applying automated video analysis techniques have been reported in various application domains such as surveillance systems [1], video retrieval [2], and human computer interaction (HCI) designs [3].

However, under the real-world settings, the large variations of video qualities and features that can be extracted are still posing great challenges to their practical usage. Another complexity can be stemmed into the ambiguity of the semantic of a so-called "event" in a video due to the subjective criteria of "event markers" that are often tailor-made by a particular application or researcher. While the changes on the illumination, colour, shape, or even textures of the studied subject over a defined period of time can be classified into the first problem domain, the extracted features from the noise-laden input can also confuse the recognition system due to inexplicit boundaries between an "event" and its "background", which naturally resides in the second problem domain.

To tackle those problems through "downgrading" the noise impact and "upgrading" the intrinsic continuous characters of video events, a so-called Spatio-Temporal Volume (STV) data structure introduced by Adelson and Bergen [4] in 1985, had been adopted in this paper. The STV represents video features in a 3D spatial and temporal volume space, which transforms the event recognition tasks into corresponding 3D feature extraction and recognition operations.

Through studying the global feature distribution, a video event can be encapsulated and defined as a specific 3D pattern that simplifies the event detection tasks on the semantic level. Global features at here mean the 3D shapes and contours composed of the segmented STV regions.

In this research, an improved pattern recognition algorithm has been developed to harness the promising characteristics of the STV structure. The core of this approach is based on the region intersection (RI) methods that compare 3D shape templates with extracted STV regions.

It has been proven that many 2D image processing and pattern recognition algorithms can be readily extended to the 3D space in a high dimensional vector form for defining advanced features such as 3D curves, shapes, volumes and their formations. The method proposed in this research is started with an optimised 3D over-segmentation operation to detect the boundaries and feature distributions of the STV sub-regions, which ignores the semantic differences between an event "actor" and its "background". This "background-independent" operation reduces the false-positive rate in differentiating noises and signals under the complex real-world settings.

After the feature extraction, a calibration mechanism has been established based on the so-called normalised RI distance. The technique evaluates the distribution of the segmented regions and then calibrates the matching distance by using a coefficient factor. Compared with original approach, this new process matches the templates

more accurately in a noisy environment.

The rest part of this paper is organised in the following order: Section 2 briefly reviews the state-of-the-art of STV and its related operations. Section 3 focuses on the algorithm design and system prototyping. The system is then tested and evaluated in Section 4. Section 5 concludes the paper with planned future works.

2. Literature review

STV was first introduced for highlighting the relationship between the temporal and pixel distributions in video footages in the 1980s [4, 5, 6]. Limited by computer performance at the time, a number of world-leading image processing research groups had attempted to map the STV models to their customised 2D projection planes before applying the normal pixel-based processing methods for further analysis. These so-called "clipping plans" inscribed with distinctive textures and shapes can be adopted to infer dynamic information, such as a human gait analysis method introduced by Niyogi and Adelson in 1994 [7]. Researchers have also developed various local spatial and temporal feature–based techniques for video event detection, such as the local grid-based model [8, 9] and the bag of interesting words (BoW) method [10, 11, 12].

Into the new Millennium, STV-based methods have gathered further momentums attributing to the increasingly powerful PC computers. Global features such as silhouettes or object boundaries [13, 14] and optical-flow [15, 16, 17] have been studied in the STV space. Wider applications of the STV-related techniques have been found in medical visualisation [18], traffic analysis [19], crime scene reconstruction [20] and crowd management [21]. These advancements have highlighted the volumetric nature of the features studied. For example, Gorelick *et al.* [22, 23] in 2007 proposed a 3D silhouettes-based shape-invariant analysis method. Through deploying the Poisson distance equation on volumetric shapes, the local spatio-temporal saliency features and Hessian orientations can be abstracted to represent the video events.

Recently, combining and calibrating the STV global features with local representations have made significant progress. For example, Jiang *et al.* [12] introduced an inter-frame constrained local feature definition for the convex matching scheme often used in human action detection. Dollár *et al.*'s [24] spatio-temporal cuboid prototyping method extended the ability of the 2D interest point into STV space. Siva and Xiang's [21] 3D BoW, Basharat *et al.*'s [25] SIFT-based video retrieval, Zhai and Shah's [26] spatio-temporal attation model, Leptev and Lindeberg's[11] slice-based feature points and Loper *et al.*'s [27] bag-of-visual-features (BoVF) have also pushed efforts on this front.

3. System development

3.1. Spatio-temporal Volume structure

This research has adopted the spatio-temporal volume (STV) data structure to represent the spatial and temporal information extracted from the original video clips and events. As illustrated in Figure 1, the STV defines a 3D volume space in a 3D coordinates system denoted by X, Y and T (time) axes. In a more observant manner, a STV model is composed of a stack of 2D arrays of pixels projecting along the orthogonal temporal axis. In this structure, the concept of an individual frame is replaced by an analogical 3D volume model in which its density, envelop and slices are all factors to the final interpretation of the model. The STV data structure transforms the video event detection process from a conventional frame-based mechanism to a 3D model analysis operation. By using this transformation, dynamic information can be represented by 3D shapes, flows or discrete points. Various pattern recognition methods, shape analysis and matching algorithms can be applied on the 3D volumetric features on an event template to solve the event detection challenges.

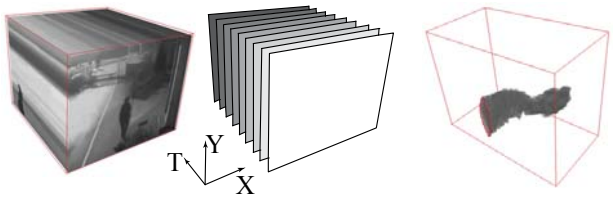

Figure 1. People falling down defined by a STV shape

3.2. Event detection system prototyping

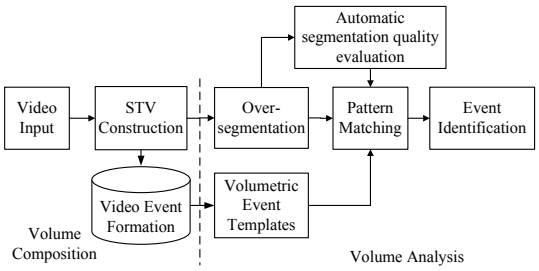

Figure 2. System pipeline

As illustrated in Figure 2, the system begins with a video signal acquisition module before data being transformed to STV models in a queuing buffer (See Section 4.2). Global segmentations then take place prior to the template matching operations being applied. The event template constriction works in this project are considered an off-line operation. Standard human actions have been defined based on popular video data bases such as KTH [28] and Weizmann [23], which include many

"elementary" single human actions and gestures such as "waving hands", "walking" and "jumping".

The baseline segmentation method used in this research is named as Pair-wise Region Comparison (PWRC), which is an accurate segmentation approach for classifying textures based on original intensity or colour features. The 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, this graph-oriented comparison approach has been applied into the STV space.

Based on the early works of Ke *et al.*'s Region Intersection (RI) method [29], an improved event detection strategy has been developed for recognising video events by matching the global features of the over-segmented regions. The matching outputs are then automatically calibrated by deploying an evaluation mechanism based on the segmentation qualities for improving the efficiency of the event recognition process. The advantage of the devised recognition method in this research has shown its distinctive advantages when dealing with real-world signals from complex backgrounds.

3.3. Feature extraction

For representing the feature point distribution of a STV model, the hierarchical PWRC segmentation algorithm is deployed [30, 31].

The PWRC is a graph-based clustering technique for representing an image's pixels and their neighbours as a graph $G=(V,E)$, where V denotes a collection of vertices v_i in the graph and E denotes the collection of edges e_i between two vertices that $(v_i, v_j) \in E, (i \neq j)$. In this algorithm, the vertices and edges are defined by corresponding regional histograms and their distances. At start, each vertex is in an independent cluster C_v. The similarity between different regions and the dissimilarity inside a region are then compared. Depending on the "similarity factor", two regions can be merged into a new larger region. During the operation, the initially independent regions will keep growing in an iterative fashion until reaching a predefined threshold.

For improving the run-time performance of PWRC in the STV domain, this research has also applied a hierarchical structure for PWRC segmentation as expressed in the following pseudo code in Table 1.

In the Table 1, the $Int(C_v)$ and $Diff(C_u,C_v)$ is calculated based on the following equations:

$$Int(C_v) = \max_{(v_i,v_j) \in MST(C,E)} w((v_i,v_j)); \quad (1)$$

$$Diff(C_1,C_2) = \min_{v_i \in C_1, v_j \in C_2, (v_i,v_j) \in E} w((v_i,v_j)); \quad (2)$$

$$MInt(C_1,C_2) = \min\left[\left(Int(C_1) + \frac{k}{|C_1|}\right), \left(Int(C_2) + \frac{k}{|C_2|}\right)\right]. \quad (3)$$

INPUTS
Spatio-Temporal Volume (STV)
Pair-wise Region Comparison factor k
Hierarchical levels n

OUTPUT
STV with Labelled segmentation regions $\mathbf{C_{n-1}}$
ALGORITHM
Transform STV colour space from RGB to L*a*b*;
Initialise $\mathbf{C}_0$ by associating the elements C_v with STV ;
Loop i from 1 to n
  Build L*a*b* histogram for each region in $\mathbf{C}_{i-1;}$

  Represent $\mathbf{C}_{i-1}$ as a graph $G=(V,E)$:
    V= L*a*b* histogram;
    E=χ^2 distance between neighbours of the histograms;

  Calculate $\mathbf{C}_i$ based on fundamental PWRC method:
    Calculate inner differences $Int(C_v)$ inside each cluster of $\mathbf{C}_{i-1}$;
    Calculate the differences $Diff(C_u,C_v)$ between two consecutive clusters of $\mathbf{C}_{i-1}$;
    Merge two cluster and renew $\mathbf{C}_i$ if $diff(C_u,C_v) < MInt(C_u,C_v)$;

  Build next hierarchical level on lower resolution
End Loop

End **Pseudo code**

Table 1. Pseudo code of Segmentation algorithm

In above equations, w denotes the weight of an edge in $Int(C_v)$. Equation 1 is defined by the Minimum Spanning Tree (MST) of the clusters; Int=0 if C_v contains only one vertex element. Since MST presents a minimum cost description of an graph, other components in the same connected graph should contain more than one edge that are larger than $Int(C_v)$ to define the lower threshold of the internal feature difference. Equation 2 represents the minimum difference between two distinctive clusters which is the lower threshold for merging two regions. The $Diff$=∞ if there is no edge connecting C_1 and C_2. Equation 3 is an improved version of the Equation 1 which introduces k for maintaining the segmentation sensitivity, where $|C_v|$ denotes the number of elements in a cluster.

The sensitivity of the PWRC segmentation is controlled by the coefficient k with larger values will leading to a bigger segmented regions. When smaller values are applied, it can ensure most of the important boundaries being extracted. This research has chosen smaller k for

extracting more accurate volumetric shapes for the following matching operations.

The PWRC processes are iterated in a range of 5 to 11-level hierarchical structures from the STV data sets. As shown in Figure 3, the video shots contain many uniformed colour areas such as ground, buildings and sky. In addition, details and texture variations such as people's clothes and windows frame of buildings are surrounded by the large solid areas. In the experiment, most of the human boundaries was identified and segmented accurately.

Figure 3. Video segmentation result by using hierarchical PWRC

3.4. Pattern recognition

As illustrated in the system pipeline, the segmentation outputs provide shape and boundaries features for representing event profiles in the STV space, which transforms the event recognition operation into a 3D shape matching operations. Various pattern matching techniques, such as [23],[13] and [14], analysing the distribution of boundary segments directly based on the assumption that the segmentation outputs contain "perfectly sorted boundaries". In reality, video events are difficult to be "clearly separated" from any uncontrolled backgrounds-against often fine-tuned background in a lab. Many "fake" regions can be falsely identified as interested regions. For example, as illustrated in the Figure 3, the "actors" in the video has been segmented into more than one part due to the texture 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. A close examination of the Figure 3 reveals that the over-segmentation has effectively identified all the intersected space actually contains all the shapes boundary sections (sub-boundaries) in the volume. The matching operations can be carried out based on these segments through analysing the distribution of them. This approach can be classified into the so-called "Region Intersection" (RI) category. One of the distinctive features of RI methods is their ability to perform shape-based event detection in challenging real-world setting where event signals are often immersed under complex background noises. The method devised in this research has explored the following design theorem.

3.4.1 Baseline of the RI Method

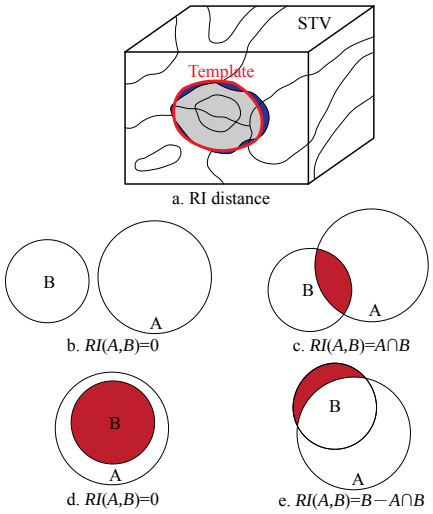

Figure 4. RI template matching algorithm based on Set Theory with four possible scenarios

Based on the Set Theory, the mechanism of a baseline RI method can be simplified as illustrated in Figure 4, where the cuboid represents an over-segmented STV. An event shape template, highlighted by the bold boundary will slide across the entire volume at runtime to identify and intersections with all 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 from Figure 4.b to 4.e. It is worth noting that in the case of Figure 4.e, where a large overlapping region is intersected with the template. The RI distance will be calculated using the Relative Complement of the template in the over-segmented sub-regions. These different scenarios can be summarized as:

$$I(E,S_w)=\begin{cases}|E\cap S_w| & if \quad |E\cap S_w|<|S_w|/2\\ |S_w-E\cap S_w| & otherwise\end{cases}, \qquad (4)$$

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:

$$I_l(E,S)=\frac{1}{N(E)}\sum_{i=1}^{n} I_l(E,S_w); \qquad (5)$$

$$N(E)=\begin{cases}\sum_{i=1}^{n}\left(\frac{|E|}{2}-\frac{|E|}{2^{|E|+1}}C_{|E|}^{|E|/2}\right), & |E|=even\\ \sum_{i=1}^{n}\left(\frac{|E|}{2}-\frac{|E|}{2^{|E|}}C_{(|E|-1)}^{(|E|-1)/2}\right), & |E|=odd\end{cases}, \qquad (6)$$

where l denotes the current location of the sliding template window. $N(E)$ is the normalised 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 template across the volume (searching window mechanism), the RI operations will mark all the locations with a matching distance less than a pre-defined threshold as a "match".

3.4.2 The Coefficient Factor-based RI Distance

The RI baseline method introduced above can detect most event corresponding shapes in an over-segmented STV. However, the accuracy of this method needs to be improved for real-world settings, where the proposed coefficient factor will improve the accuracy of the RI distances. As discussed in Section 3.3, real-world video inputs usually contain both large solid colour areas and detailed textures. The PWRC segmentation methods can classify these contents in an over-segmented style consisted of both large and small sub-regions. But some extremely smaller regions around the event shape boundaries can produce substantial normalised RI distances, which is a potential cause for the intolerance to the false negative results. In the new approach, "rewards" have been given by to the larger sub-regions that effectively reduce the distance values. For the smaller regions, contrary measures will be taken automatically.

The evaluation scheme is automatically generated based on a quantifying process to the intersected regions' local histogram to record the size and the number of the intersected sub-regions as shown in Figure 6.

The local histograms discussed above can be used to compare with the histograms extracted from the controlled ideal RI matching scenarios. When the event actors and backgrounds are perfectly segmented as illustrated in Figure 5, 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 different situations when calculating the coefficient factors as illustrated in Figure 6.a, b and c, where the event templates are denoted by the artificial ellipses. In Figure 6.a, 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 Figure 6.b, the histogram is showing a largely flat curve with small fluctuations indicating a fewer but larger intersected regional blocks. Figure 6.c contains both large and small intersectional parts, where the smaller regions are dominant; 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 normalised χ^2 distance between current histogram and "perfect" histogram.

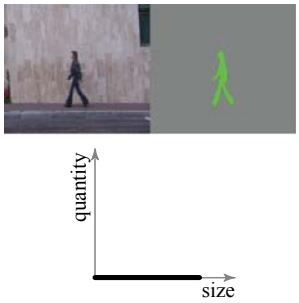

Figure 5. "Perfect" segmentation defined by event template

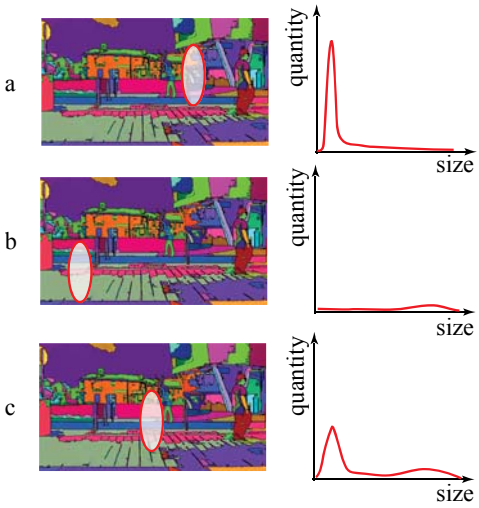

Figure 6. Local histogram used for evaluating the intersected segmented regions

The coefficient factor can be expressed as a linear transformation from the histogram distance as denoted in Equation 7:

$$\tilde{I}_l(E,S)=I_l(E,S)\cdot[a+b\cdot f_l(E,S)], \qquad (7)$$

where $f_l(E,S)$ is the normalised 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 change is in

between 0.6 and 1.4, which is proven suitable for most of the video datasets testes.

4. Tests and evaluations

A series of experiments have been designed and carried out for validating the system performances under the controlled and real-world settings. The system software was developed by using the corresponding MATLAB functions, the OpenCV 2.2 libraries and LabVIEW API (Application Program Interface). The experiments were mainly carried out on a host PC equipped with an AMD 2.62 GHz Athlon CPU with 2G RAM.

4.1. System performance on public datasets

Experiments were carried out in this research to assess the event detection accuracy based on the theoretical structure of the system as introduced in Section 3.

The design theorem of these tests was to establish the relationships of the ground truths based on the calculated matching coefficients from the KTH [28] datasets. Table 2 lists the values of all parameters used in the experiment. In the KTH dataset, the event templates applied for RI matching were defined by averaging four volumetric contours in each event category.

PWRC		Coefficient Factor	
k	*n*	*a*	*b*
370	7	0.7	0.7

Table 2. Parameters used for testing public datasets

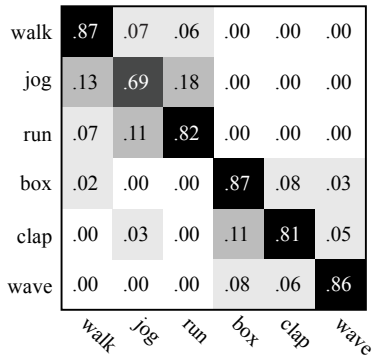

Figure 7. The KTH confusion matrix

Methods and techniques	Event Detection Accuracy
This Method: PWRC + RI + CF(Coefficient factor)	**82.0%**
Ke *et al.*'s MS (Mean shift) + RI + Flow [29]	80.9%
Schuldt *et al.* [28]	71.7%
Dollár *et al.* [24]	81.2%
Niebles *et al.* [32]	81.5%

Table 3. The Accuracy performance compared with other approaches.

Figure 7 shows the test results of the detection accuracy based on the confusion matrix acquired from each dataset. The average accuracy of developed the system is 82.0%, which is better than many popular methods as listed in Table 3. As Illustrated in the confusion matrix, certain events such as jog-and-run and box-and-clap groups are difficult to distinguish due to their silhouette similarities and small variation on the temporal axis. One possible solution to such a problem is to combing the machine learning algorithms with the local spatio-temporal features for differentiating the details of human gestures.

4.2. Performance under uncontrolled and complex backgrounds

Above experiments were carried out in controlled environments for proof-of-idea. The experiments introduced in this section were focused on the real-world performance of the system subjecting to noise and other more challenging real application settings. The parameters adopted for these testes are as defined in Table 4

PWRC		Coefficient Factor	
k	*N*	*a*	*b*
500	11	0.6	0.8

Table 4. Parameters applied for uncontrolled testing

Figure 8. Footages (bending, waving two hands and running) used for uncontrolled and complex backgrounds testing

Since the Weizmann database [23] mostly contains simple backgrounds and a single actor in each video clip, these files are ideal for defining human actions by using active contour-based [33] segmentation processes and being facilitated by manual adjustment. In this experiment, the entire Weizmann database was used for template composition through averaging the extracted template STV shapes from the same event category.

Several footages recorded in the university campus have also been tested as shown in Figure 8, which contains various action events for comparing with the predefined templates and classifiers. The length of each video is 10 minutes. The system's robustness has then been validated by the Receiver Operator Characteristic curves (ROC curves).

For ensuring the efficiency of the proposed system, this research has introduced a queuing buffer for live video

feeds, in which the incoming frames are transformed into the 3D STV space before being tested against the pre-defined templates. At runtime, the system starts with building a buffer (assigning memory) to the incoming video stream. The index of the first frame and the size of a STV model are customisable and dependant on the pre-defined action templates in terms of video configuration data such as resolution and frame rate. Once the RI-matching is completed, the buffer assigned for holding the STV model will be freed from the memory to avoid accumulating memory footprints for the next STV operations.

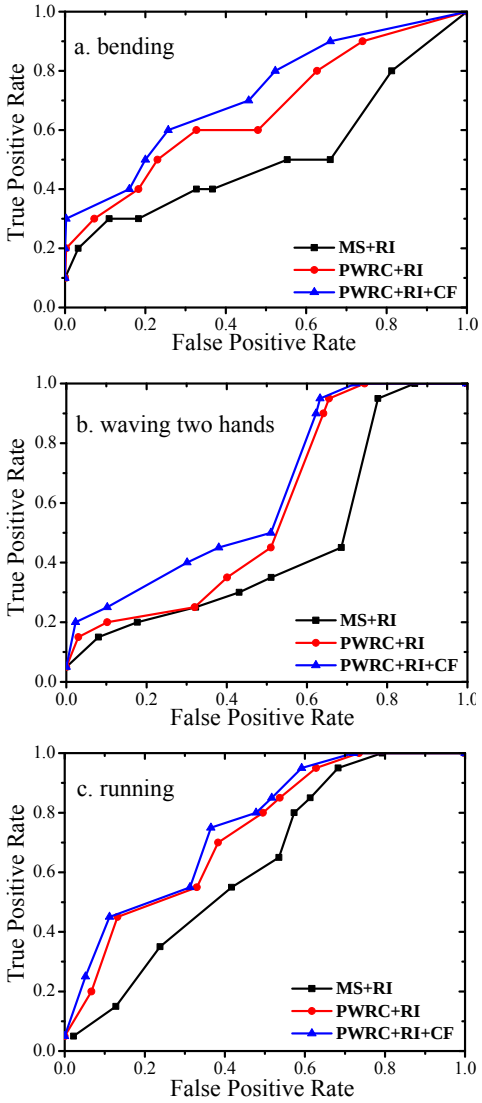

Figure 9. The ROC curves under complex backgrounds

The average time consumption of event detection processes is around 30 seconds by using parameters listed in Table 4, which includes 10s for video segmentation and 20s for template matching. The system has shown its comparable efficiency against many similar approaches such as [29] and [13].

Figure 9 demonstrated the ROC curves generated from experiment to highlight the performance difference between the RI-based matching algorithms during the changing of detection threshold (increase 10% for each plot in the curves). It is evident that the system can produce superior performances through introducing the coefficient factor mechanism as explain in Section 3.4.2. In addition, the innovative Hierarchical PWRC method can abstract more accurate shape features in compression with other clustering based segmentation methods, such as Mean Shift.

5. Conclusions and future work

This project has developed a STV-based approach to tackle the common problems in video event detection, where video contents can be modelled as 3D volumetric shapes for template matching. One of the key techniques developed in the research is the improved Region Intersection-based shape matching, which can handle the shape feature generated from the over-segmentation process. In the system design, the distributions of the over-segmented regions have also been used for the shape matching in the forms of coefficient factors, which are indicated by the relevant histograms and being evaluated as the distance matching at run-time.

The research system of its current form can only handle video events that possess distinctive shapes changes. It cannot register motions occurred inside of a volume, for example, a front view of a human hand-clapping event. It is envisaged that the future works of this project will focus on integrating other analytical features in the volume space to yield more intrinsic and non-intuitive information for interpretation those information. Currently the STV-based matching operation is still the dominant and most time consuming process in the process pipeline. Although this research was not initially aiming at real-time applications but to investigate an interactive and off-line analytical tool, the operational efficiency still plays a vital role in its future success. One of the envisaged solutions is to employ hardware acceleration by adopting the latest computer parallel architectures, for example, through harnessing data parallelism embedded in the modern Graphics Processing Units (GPUs) to facilitate the data intensive and filter-driven computations [34].

References

[1]. M. Valera and S.A. Velastin. Intelligent distributed surveillance systems: a review. IEE Proceedings Vision, Image and Signal Processing, 152(2):192-204,2005.

[2]. H.J. Zhang, J. Wu, D. Zhong and S.W. Smoliar. An integrated system for content-based video retrieval and browsing. Pattern Recognition, 30(4):643-658,1997.
[3]. J.M. Rehg and T. Kanade, DigitEyes: vision-based hand tracking for human-computer interaction, in Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects. 16-22,1994.
[4]. E.H. Adelson and J.R. Bergen. Spatiotemporal Energy Models for the Perception of Motion. Journal of the Optical Society of America A, 2(2):284-299,1985.
[5]. H.H. Baker and R.C. Bolles. Generalizing Epipolar-Plane Image Analysis on the spatiotemporal surface. International Journal of Computer Vision, 3(1):33-49,1989.
[6]. Y. Ricquebourg and P. Bouthemy. Real-Time Tracking of Moving Persons by Exploiting Spatio-Temporal Image Slices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):797-808,2000.
[7]. S.A. Niyogi and E.H. Adelson, Analyzing and recognizing walking figures in XYT, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 469-474,1994.
[8]. S. Mitra and T. Acharya. Gesture Recognition: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37(3):311-324,2007.
[9]. I. Laptev and P. Perez, Retrieving actions in movies, in IEEE 11th International Conference on Computer Vision, 2007. 1-8,2007.
[10]. J. Shi and C. Tomasi, Good Features to Track, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 593-600,1994.
[11]. I. Laptev and T. Lindeberg, Space-time interest points, in 9th IEEE International Conference onComputer Vision, 2003. Proceedings. 432-439 vol.1,2003.
[12]. H. Jiang, M.S. Drew and Z.N. Li, Successive Convex Matching for Action Detection, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, . 1646-1653,2006.
[13]. Y. Alper and S. Mubarak. Actions sketch: a novel action representation. in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.2005.
[14]. D. Weinland, R. Ronfard and E. Boyer. Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding, 104(2-3):249-257,2006.
[15]. P. Matikainen, M. Hebert and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. in 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops).2009.
[16]. D.J. Fleet, M.J. Black, Y. Yacoob and A.D. Jepson. Design and Use of Linear Models for Image Motion Analysis. International Journal of Computer Vision, 36(3):171-193,2000.
[17]. D.M. Gavrila. The Visual Analysis of Human Movement: A Survey. Computer Vision and Image Understanding, 73(1):82-98,1999.
[18]. L. Gao, D.G. Heath, B.S. Kuszyk and E.K. Fishman. Automatic Liver Segmentation Technique for Three-dimensional Visualization ofCT Data. Radiology, 201:359-364,1996.
[19]. N. Mitarai and H. Nakanishi. Spatiotemporal Structure of Traffic Flow in a System with an Open Boundary. Physical Review Letters, 85(8):1766,2000.
[20]. L.A. Nelson and S.D. Michael. The application of volume deformation to three-dimensional facial reconstruction: A comparison with previous techniques. Forensic Science International, 94(3):167-181,1998.
[21]. P. Siva and T. Xiang, Action Detection in Crowd, in British Machine Vision Conference 2010, F. Labrosse, et al., Editors. 9.1-9.11,2010.
[22]. L. Gorelick, M. Galun, E. Sharon, R. Basri and A. Brandt. Shape Representation and Classification Using the Poisson Equation. IEEE Trans. Pattern Anal. Mach. Intell., 28(12):1991-2005,2006.
[23]. L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri. Actions as Space-Time Shapes. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(12):2247-2253,2007.
[24]. P. Dollár, V. Rabaud, G. Cottrell and 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. 65-72,2005.
[25]. A. Basharat, Y. Zhai and M. Shah. Content based video matching using spatiotemporal volumes. Computer Vision and Image Understanding, 110(3):360-377,2008.
[26]. Y. Zhai and M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in Proceedings of the 14th annual ACM international conference on Multimedia,2006.
[27]. A.P.B. Lopes, R.S. Sloveira, J.M.d. Almeidq and 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. 315-321,2009.
[28]. C. Schuldt, I. Laptev and B. Caputo, Recognising Human Actions: A Local SVM Approach, in International Conference on Pattern Recognition. 32-36,2004.
[29]. Y. Ke, R. Sukthankar and M. Hebert. Volumetric Features for Video Event Detection International Journal of Computer Vision, 88(3):339-362,2010.
[30]. P.F. Felzenszwalb and D.P. Huttenlocher. Efficient Graph-Based Image Segmentation. International Journal of Computer Vision, 59(2):167-181,2004.
[31]. M. Grundmann, V. Kwatra, M. Han and I. Essa, Efficient Hierarchical Graph-Based Video Segmentation, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2010,2010.
[32]. J.C. Niebles, H. Wang and F. Li. Unsupervised Learning of Human Action Categories Using Spatial-TemporalWords. International Journal of Computer Vision, 79(3):299-318,2008.
[33]. T.F. Chan and L.A. Vese. Active contours without edges. IEEE Transactions on Image Processing, 10(2):266-277,2001.
[34]. R. Ellis, T. Peters, A. Lefohn, J. Cates and R. Whitaker, Interactive, GPU-Based Level Sets for 3D Segmentation, in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003. 2003, Springer Berlin / Heidelberg. p. 564-572.