Wang, Jing and Xu, Zhijie

Video Analysis Based on Volumetric Event Detection

Original Citation

Wang, Jing and Xu, Zhijie (2010) Video Analysis Based on Volumetric Event Detection. International Journal of Automation and Computing, 7 (3). pp. 365-371. ISSN 1476-8186

Video Analysis Based on Volumetric Event Detection

Jing Wang Zhi-Jie Xu*

Department of Informatics, School of Computing and Engineering, University of Huddersfield, Huddersfield, HD1 3DH, UK

Abstract: During the past decade, the feature extraction and the knowledge acquisition based on video analysis have been extensively researched and tested on many applications such as Closed-Circuit Television (CCTV) data analysis, large-scale public event control, and other daily security monitoring and surveillance operations with various degrees of success. However, since the actual video process is a multi-phased one and encompasses extensive theories and techniques ranging from fundamental image processing, computational geometry and graphics, machine vision, to advanced artificial intelligence, pattern analysis, and even cognitive science, there are still many important problems to resolve before it can be widely applied. Among them, the video event identification and detection are the two prominent ones. Comparing with the most popular frame-to-frame processing mode of most of today's approaches and systems, this project reorganizes the video data as a 3D volume structure which provides the hybrid spatial and temporal information in a unified space. This paper reports an innovative technique to transform original video frames to 3D volume structures denoted by spatial and temporal features. It then moves on to highlight the volume array structure in a so called "pre-suspicion" mechanism for later process. The focus of this report is the development of an effective and efficient voxel-based segmentation technique suitable to the volumetric nature of video events and is ready for deployment in 3D clustering operations. The paper concludes at the performance evaluation of the devised technique with further discussions on the future work for accelerating the pre-processing of the original video data.

Keywords: Spatio-temporal volume, video processing, volume feature extraction, segmentation; motion analysis.

1 Introduction

Inherited from image processing techniques, traditional video event detection approaches put more emphasis on spatial signal features through the frame-by-frame (FBF) processing methods [1]. However the FBF mechanism results in the loss of unabridged dynamic information contained in a video. This insufficiency leads to high false positive rate during the event detection. Generally speaking, an event in a video can be defined by correlating the coordinates of a group of related pixels through a set of frames dispersed along the temporal axis. Unlike the features extracted from a static image, a video event can record dynamic "actions". More specifically, a video event is the collection of "changes" occurred in a Euclidean space over a period of time elapsed. Both the recorded spatial and temporal signals can either be continuous or discrete. At the level of information systems, multiple events can contribute to the generation of "knowledge" that can be handled by machine intelligence or human intervention. For example, a video footage of a football match can contain many events such as tackling, jumping, and running.

The definition of video events introduced above has brought in the concept of time elapsed in video processing. This research adopts the spatio-temporal volume (STV) data structure to represent spatial and temporal features from original video clips. As shown in Fig.1, the STV defines a 3D volume space in a 3D coordinate system denoted by X, Y and T (time-dimension) axes. In a more natural point of view, it is composed of a stack of video frames formed by arrays of pixels in the time order. In this structure, individual frame is represented by the mappings of the X-Y coordinates with the corresponding pixel values, while the dynamic information of the events is largely maintained through the navigation along the time axis. To integrate the spatial (coordinates) and temporal (time) information in a single data structure, each fundamental element inside of the STV "box" is called a voxel-acronym of volume-pixel, which conjoins the pixel and the time information together.

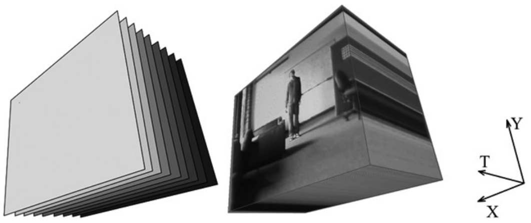

Fig. 1 STV Structure

The STV data structure transforms video event detection approaches from a FBF-based mechanism to one of the 3D-oriented shape analysis operations. Useful events can then be extracted directly from the volume data by deploying appropriate feature matching processes, which are mainly rely on 3D segmentations and clustering stemmed from conventional image processing techniques. As shown in Fig.2, a video event of a "waving hand" has been extracted to form a STV model. It demonstrates the feature segmentation operation for highlighting the contour of non-rigid body changes through denoting the original STV into the potential features and the background.

If different video events can be abstracted and modeled

Manuscript received date; revised date

*Corresponding author. E-mail address: z.xu@hud.ac.uk

as 3D templates-shapes, then the corresponding event detection tasks can be transformed into the jobs of recognizing 3D shapes from input video volumes. In practice, a 3D template shape can sometime show an event in the form of the contour of a subject, but more often, a 3D shape is marked by a group of voxels that are not visually comprehensible, such as the trajectories of some discrete points.

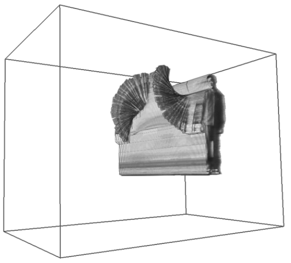

Fig. 2 STV model of a waving hand

This project has two main objectives. The first one focuses on the volume feature extraction, which forms the core of this paper. The second attempts to solve some difficult feature analysis problems, which will be reported in a separate article. The paper is organized in the following order: Section 2 provides a brief review on the existing STV analysis techniques. Section 3 introduces the proposed STV shape modeling method based on the so called "STV-array" which segments the STV cube along the time axis; Section 4 highlights the related 3D voxel-based segmentation techniques devised in this research with experimental results. The result is further analyzed and discussed in Section 5 with Section 6 concludes the work.

2 Literature review

The volume data structure mentioned earlier emphasizes the temporal continuity in an input stream of a video data. The use of spatio-temporal volumes was first introduce in 1985 by Aldelson and Bergen [2], who build motion models based on "image intensity energy" and the impulse response to various filters. There are a number of widely deployed methods for analyzing the STV. One of them is through slicing a stack of two-dimensional temporal slices for dealing with a variety of problems. For examples, inferring feature depth information [3], generating dense displacement fields [4], camera calibration [5], motion categorization [6], tracking [7], ego-motion estimation [8]; as well as in many application system such as advanced navigation [9] and view synthesis systems [10].

For the particular application of event detection, the most popular 3D volume-based approaches are the so called shape-based methods. For example, all the human gestures can be modeled as non-rigid action templates for automated sign language interpretation. The success of this kind of shape-based analysis relies heavily on the quality of the segmentation process. If deployed successfully, the shapes or the contours of the shape will yield significant features which can be used for possible events occurred.

Comparing the aforementioned 2D slice-based process, the 3D-based approaches can reveal more hidden features if appropriate segmentation operation are applied. For instance, a volume can show a series human contour that accumulates the 3D shape of a human silhouette. Therefore, the aim of the volume shape-based human event detection is to evaluate the 3D STV with enriched shape information to facilitate the investigation of the types of event is occurred over the time span.

Shape-based methods generally employ a variety of techniques to characterize the shape of an event, for example, shape invariants [11-13]. To improving the computational efficiency and robustness of the extracted action variations, Lena [14] introduced a method to analysis 2D shapes to through integrating information introduced by human behaviors. This method applies the Poisson equation for extracting various shape properties that are utilized for shape representation and classification.

Bobick and Davis [15] have used the spatio-temporal volume for generating motion-history images, which was extended by Weinland et al. [16] for handling motion history volumes, which is more practical and flexible to implement. It is simple to operate on due to its time information has been regarded as an additional dimension from a 2D motion history image (the different intensity of the pixels means the different time sequence). In its data structure, the changes time over are reflected by the gradual pixels intensity changes. The direction and speed of the motion can then be easily represented in a single 2D image, where the optical flow-like motion vectors can be calculated from the gradient of the motion history image directly [17].

3 STV construction and STV array structure

STV is a 3D volume data structure, which is widely used in medical visualizations, such as Magnetic Resonance Imaging (MRI) scan [18]. This project has chosen the STV to define and detect events in videos. As a pre-processing step, it is necessary to change the original digital video format to the STV.

3.1 Digital video conversion

Conversional digital video is an aggregation of 2D frames in time order. Each frame shares the same size unless redefined, which can be expressed as:

$$V = \{F_1, F_2, \ldots, F_n\}, \tag{1}$$

where V denotes a specific video file and F_i (i=1, 2, … , n) denotes individual frames of the video, where the n indicates the total frame number of the video. Each frame is identical as in the image plane $D \subset \mathbf{R}^2$. A point $\mathbf{p} \in D$ is referred as a pixel. Considering the simpler case of gray scale for the image plane, each pixel can be represented as $\mathbf{p}$=(p_i,p_k) where j and k denote the coordinate values of the pixel in the 2D image plane. The function I=I($\mathbf{p}$) preserves the pixel value, in gray scale.

In contrast, the STV structure preserves the video information through the use of voxels, where

$$\mathbf{v} \in \mathbf{R}^3, f(\mathbf{v}) \in \mathbf{R}. \tag{2}$$

The $\mathbf{v}$=(v_x,v_y,v_z) indicates a voxel in 3D space. The

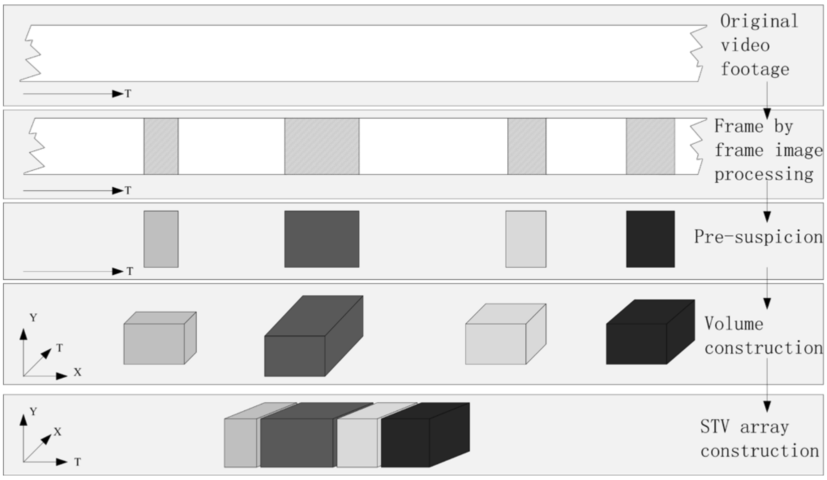

Fig. 3 Pre-suspicion STV array mechanism

function *f* preserves the voxel value, in gray scale. This research stores the 3D matrix into a 1D array in the "front-left-top" and "right-down-backwards" style, where the direct volume rendering (DVR) techniques are used for result visualization.

3.2 STV array structure for efficient voxel-based processing

Video data from real applications usually contains thousands of frames. It is both unnecessary and impossible to compose and analyze all the video events in a single enormous STV structure. A conceivable solution is the adoption of the STV array structure based on a pre-processing mechanism which decomposes the video into useful and useless "paragraphs". It constructs a series of sub-STVs by marking interesting features in each frame. This mechanism rebuilds a quantity of "pre-suspicion" STV data from original video footage and composes the STV sequence as an STV array.

Pre-processing steps can remove many frames which make little contribution for event analysis in the original video. As shown in Fig.3, the residual video clips are translated into "pre-suspicion" STV structures which have a high probability of containing events. Since the video footage contains many events, pre-suspicion adds to the number of STV, but reduce the complexity of analysis. Each STV in the array might only contain one event depending on the definition of it. The complexity of the follow up steps such as segmentation and pattern recognition can then be simplified. The FBF processing for the "pre-suspicion" STV also provides useful 2D features which can also be used in event detection. For example, based on the edge information, shadows from traffic sequences are successfully removed by Xiao [19]. Yamamoto also introduced a colour detection approach for multi-view video by using energy minimization of view network [20].

3.3 Video volume compression

As explained in Section 3.1, original STV data catch the video frames one by one according to the temporal order and convert them into 2D slices to form a 3D stack. This process preserves every pixel in a video and transforms them into the 3D space as voxel. This process order can introduce significant size problem. For example, 5-second video clip at a frame rate of 30 with the resolution of 320 by 240 pixels will result at 33MB memory consumption at run time for just viewing the data block before any further process.

To tackle this problem, a new feature-based volume structure has been developed in this research which consists of two main parts, the frame pre-processing and the volume compressor. The prior will filter the original frames and only keep the "useful" features in each frame, which means before the 3D volume is through applying various traditional image processing techniques such as optical flow [21] and the partner recognition approaches. This method removes the large still background pixels and separates the useful features according to specific application. This pre-processing step ensures a low level of entropy through constructing a feature-only STV volume.

As appropriate compressing technologies can further reduce the memory footprint, the latter part of the devised process applies an Audio Video Interleave (AVI) compression filter to produce the final STV feature volume. Other popular compression techniques and file structures might be used for this purpose too, such as the Moving Picture Experts Group (MPEG). Applied on the case addressed earlier in this section, the 5-seconds video clip at a 30 fps and in resolution of 320 by 240 will only 130KB in the memory if stored as an AVI file in the Digital Video Express (DVIX) code.

4 Voxel-based segmentation by clustering

The segmentation process divides a volume into constituent sub-regions. The level to which the subdivision is carried out depends on the problem being solved, which means the segmentation process should cease when the regions of interest in an application have been isolated.

The STV segmentation methods devised in this research so far are mainly based on extending the 2D image segmentation techniques into 3-Domain. In the 3D environment, the volume segmentation process is similar to sculpturing in which unnecessary parts of a raw block are removed from the bulk. For a STV "cube", the "things" to be removed can be defined by various features such as colour, density, edge and texture [22]. As shown in the Fig.2, this volume of waving event has been segmented by isolating the active contour. After volume segmentation, a representing 3D feature volume in the feature space can be built for further event recognition task. In this research, the clustering approaches are employed due to their efficiency and robustness.

The clustering methods in general intend to sort the studied elements using the pre-defined spectrums. In the volume space, voxel groups are categorized by different signatures. The volume segmentation process can benefit from 2D-based methods such as K-Mean and Mean-Shift clustering approaches without changes on the foundational mathematic model. The only difference from the pixel-based operations is the extra dimension in the 3D feature space.

Taking the Mean-Shift (MS) clustering algorithm as an example, the original MS method was presented by Fukunaga and Hostetler [23] as a nonparametric method to estimate a Probability Density Function (PDF) using the so-called Parzen window density estimator [24]. Using a similar notation as explained in [25], the MS technique can be described as follows: given n data points $\mathbf{x}_i$, i=1,...,n in the d-dimensional (R^d) feature space, the multivariate kernel density estimator with kernel $K_\mathbf{H}(\mathbf{x})$ computed at the point $\mathbf{x}$ is given by

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_H(\mathbf{x}-\mathbf{x}_i), \tag{3}$$

where H is a symmetric positive $d \times d$ bandwidth matrix. $K_\mathbf{H}(\mathbf{x})$ can be defined as

$$K_\mathbf{H}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x}). \tag{4}$$

The multivariate kernel function $K(\mathbf{x})$ is a bounded function which obeys following rules.

$$\begin{gathered} \int_{R^d} K(\mathbf{x})d\mathbf{x} = 1 \qquad \lim_{\|\mathbf{x}\|\to\infty}\|\mathbf{x}\|^d K(\mathbf{x}) = 0 \\ \int_{R^d} \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0 \quad \int_{R^d} \mathbf{x}\mathbf{x}^\mathrm{T} K(\mathbf{x})d\mathbf{x} = c_K\mathbf{I}, \end{gathered} \tag{5}$$

where C_K is a constant. Two well-known Kernel functions, $K^P(\mathbf{x})$ and $K^S(\mathbf{x})$ can be generated from a symmetric kernel $K_1(\mathbf{x})$.

$$K^P(\mathbf{x}) = \prod_{i=1}^{d} K_1(x_i) \tag{6}$$

and

$$K^S(\mathbf{x}) = a_{k,d}K_1(\|\mathbf{x}\|), \tag{7}$$

where $K^P(\mathbf{x})$ is generated by product of the kernel and $K^S(\mathbf{x})$ is obtained from rotating $K_1(\mathbf{x})$.in a specified manner. The $a_{k,d}$ is a normalization constant.

In this project, only a special class of radially symmetric kernels that predigest (7) for $x \geq 0$ is of interest to the investigators. The normalization constant is assumed strictly positive in this case as shown in (8.1).

$$K(\mathbf{x}) = c_{k,d}k(\|\mathbf{x}\|^2), \tag{8.1}$$

where profile k is defined by a gate function

$$k(x) = \begin{cases} 1-x & 0 \leq x \leq 1 \\ 0 & x > 1 \end{cases} \tag{8.2}$$

It also predigests $\mathbf{H}$ from fully parameterized matrix to an identity matrix $\mathbf{H}$=$h^2\mathbf{I}$, where h is the window size of the Mean-Shift. It is clear that the latter case is the only bandwidth parameter need to be provided before Mean-Shift operation.

After introducing (7) and (8.1) into the kernel density estimator (3), the proximate expression of (3) is

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right), \tag{9}$$

which quality can be measured by the mean of the square errors between the densities and their integrated over the domain.

The Mean-Shift operation finds the peak value in the feature space and then classifies relevant feature points in the nearby area. In the density estimator model, different peak values can belong to different maximum density areas, which mean the modes are located among the zeros of the gradient $\nabla f(\mathbf{x}) = 0$. This can be explained in the following expression:

$$\nabla\hat{f}_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(\mathbf{x}-\mathbf{x}_i)k'\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right). \tag{10}$$

and

$$g(\mathbf{x}) = -k'(\mathbf{x}), \tag{11}$$

where k' is the derivative of the profile k. Therefore transform the formula (10) into

$$\begin{aligned} \nabla\hat{f}_{h,K}(\mathbf{x}) &= \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{x})g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right) \\ &= \frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)\right] m_{h,G}(\mathbf{x}). \end{aligned} \tag{12}$$

As proven by Comaniciu et al. [25] that the Mean-Shift vector $m_{h,G}(\mathbf{x})$ can be expressed as:

$$m_{h,G}(\mathbf{x}) = \frac{h^2c}{2}\frac{\nabla\hat{f}_{h,K}(\mathbf{x})}{\hat{f}_{h,G}(\mathbf{x})}, \tag{13}$$

where K and G are kernels with respective profiles k and g.

where h is the bandwidth of the kernel used, and c is the normalization constant. (13) indicates that the Mean-Shift vector is aligned with the local gradient estimate, hence it can be used to detect the local maxima of this distribution [26]. The main difference between MS and other nonlinear clustering methods is how the information in the spatial and range domains is treated to obtain the filtered image. Basically, MS can be seen as an adaptive gradient ascendant method.

For 2D image processing, it is usually referred to the space coordinates and the colour value of the 2D pixels in the feature space. Consequently, the feature space generated is a 5D space (x,y,r,g,b), in which (x,y) denotes the space coordinates and (r,g,b) the colour of the pixel. These five elements represent a single point $\mathbf{x}_i$ in the feature space. After all pixels are mapped, the multivariate kernel density estimator developed by Duda and Hart [24] can be deployed for the MS arithmetic.

In the case of STV, this analytical mechanism can still be applied, but the pixel will be replaced by voxel as studied element. The feature space will become a 6D space define as (x,y,t,r,g,b), where (x,y,t) denotes the space coordinates and (r,g,b) the color of voxels. The identical multivariate kernel density estimation can then follow suit.

5 Experiment result

To assess the devised STV feature model and the corresponding segmentation and clustering operations, a set of experiments have been designed and carried out to evaluate the system performances. The software tools and Application Programming Interface (API) used in those experiments include, MATLAB, LabVIEW, OpenCV, OpenQVis and the system prototype is implemented in VC++ on a AMD Athlon 2.62GHz GPU with 2G RAM.

5.1 STV array structure

A short video clip was captured using NI 1411 image acquisition card connected to a colour CCTV, with a frame rate at 10 fps and a frame size of 640 by 480. This experiment defines the pre-suspicion mechanism through restricting the interested region only on moving objects in the video. It then composes STV shapes by assembling the moving contours. As mentioned in Section 3, these steps must be completed before each STV cube is established. The output should contain a series of STV arrays and each small STV element should contain only one event with non-rigid moving contour. Fig.4 shows this algorithm in a state transition diagram. In this algorithm, the moving object is abstracted directly through removing the static background. This background was identified by the “median background” technique [27] and was calculated by capturing 100 frames on a 100ms separation.

Following the “capture a frame” operation, a FBF-based image processing was performed to find the canny edge [28] for the contour in the absolute- difference image. The high and low-threshold is set at 70% and 30% of the maximum pixel value. The size of the Gaussian smoothing filter was 9 by 9, and the result of which is shown in the Fig.5.

Another important feature in this algorithm is the logic flags which control the state transition and are marked as “Pre-suspicion?” and “Operation Completion?” in Fig.4. After concluding on whether the current frame contains moving contours in the “Pre-suspicion” phase, different process combinations of these flags will lead to different transition directions.

5.2 Voxel-based segmentation

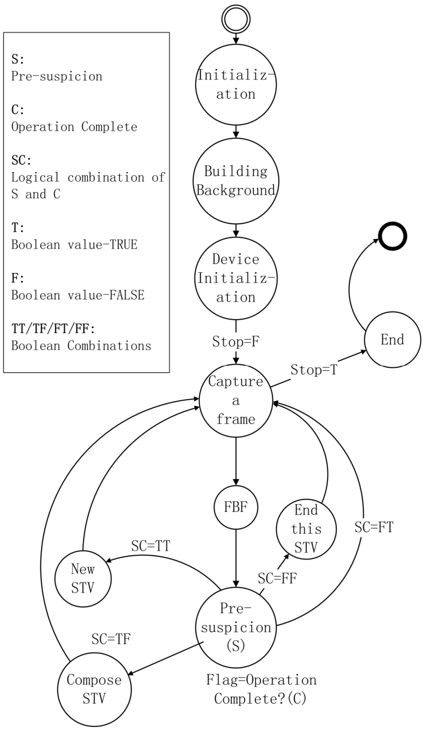

Fig. 4 Pre-suspicion algorithm

This research had initially focused on evaluating the K-Mean and MS clustering operations on the STV segmentation. Three STV volumes were constructed for this purpose - “Waving one hand”, “Waving both hands” and “Walking” events - as shown in Fig.6.

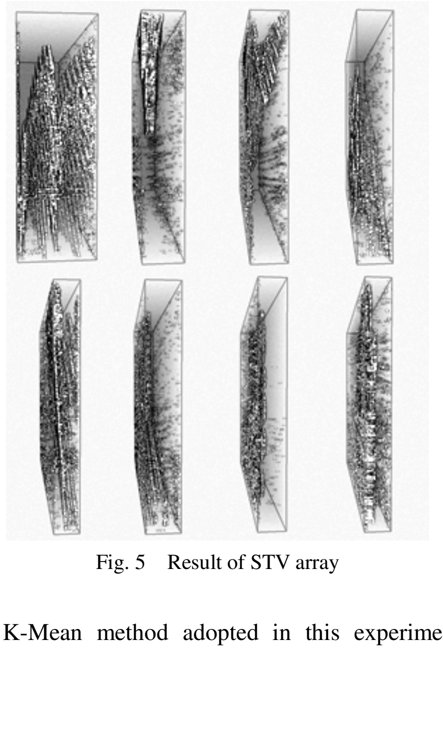

Fig. 5 Result of STV array

The K-Mean method adopted in this experiment is based on the intensity of the gray-level for each voxel. The approach is a natural extension of the 2D-based pixel operations since the only difference is the extra dimension introduced by the voxel. It can be readily handled by the vector expressions of most of the classic clustering algorithms. The devised 3D algorithm can be explained in the following pseudo code.

```
1.  Define the value of $k=n$. This value $n$ is the number
    of clusters.
2.  Initialize the distribution of each cluster. These
    centers of clusters should be distributed equally in
    the 3D feature space.
3.  While (center of each cluster is still changing)
    a)  Compare each feature point with all the centers
        of cluster.
    b)  Put the feature point p into cluster C$i$ if the
        distance between p and center of Ci is the
        shortest.
    c)  Remove the center to a new position if new
        element p is put into this cluster.
    End While
4.  Show the clustering result
    End of the algorithm
```

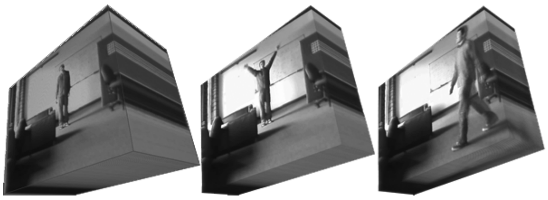

Fig. 6 Original STV contains different events

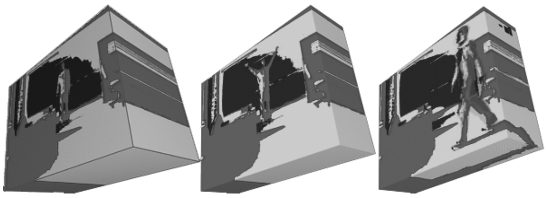

Fig. 7 3D K-Mean Segmentation result

The actual experiment result is illustrated in Fig.7. The time consumption of the operations is shown in Fig.8. It is clear that even for the relative simple operations such as K-Mean to be applied on the 3D volume space, the average time cost is substantial. Some anticipated solutions for alleviating this problem will be discussed in the last section.

The Mean-Shift (MS) clustering technique has also been experimented in this project. As discussed in Section 4, the voxel-based MS will extend the feature space from 5D to 6D. The MS algorithm developed in this experiment is based on Dorin Comaniciu and Peter Meer`s work [25]. The results can be seen in Fig.9.A, 9.B and 9.C (with both the Hr and Hs set at 32) representing the waving one hand, waning both hands and walking events, respectively.

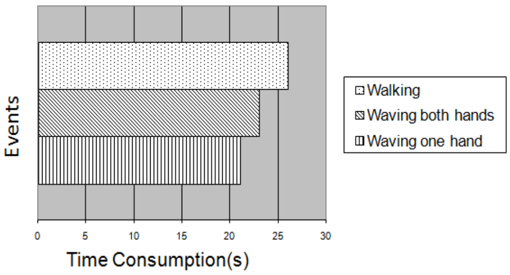

Fig. 8 Time consumption of voxel-based K-Mean segmentation

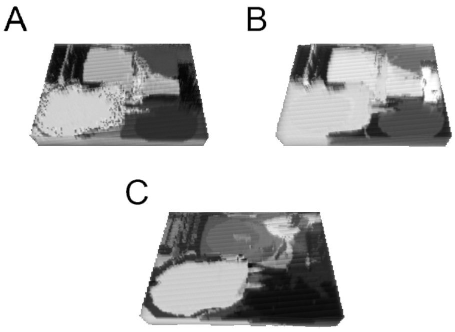

Fig. 9 Voxel-based MS operation

6 Conclusion and future works

The main research aim of this project is to realize video volume-based event detection and to investigate the relevant key techniques, which have led the design and development of a general framework of the study. The investigation can be divided into two main phases: 3D segmentation and 3D template mapping. The work reported in this paper has focused on the prior, in which the main contribution is a clear guideline for extracting 3D features from the volumetric data structure.

This project has introduced the STV structure for handling video contents and the construction of a pre-suspicion STV array to introduce an efficient way to analyze the STV event shapes. Based on established 2D image processing techniques such as the K-Means and MS, a number of clustering and segmentation techniques have been successfully transformed into the 3Dvolume space. One of the anticipated key tasks in the future is to devise an efficient template mapping technique which can be applied to the pre-built STV feature volumes for event identification in large digital video repositories.

As evident in the experiments detailed in Section 5.2, the complex volume data structure has introduced substantial time-consumption when processing the STV. It is well known that the K-Mean method is an efficient segmentation technique in 2D image processing. However, when applied into 3D domain, the performance deteriorated rapidly. For more complex operations, such as the Mean-Shift, this contains many iterative steps, the run time of the algorithmic steps become even more intolerable. One of the potential solutions for solving this problem is through hardware acceleration, for example, to employ the Graphics Processing Unit (GPU) for accelerating the computation [29]. It is understood in this research that most STV processing techniques handle each voxel using the same arithmetic operations, which can be realized in programmable GPU streams in the SIMD (Single Instruction Multiple Data) processing mode. The acceleration factor has been proven in many early studies. For example, comparing to the CPU-dominant approach, the Meer's [30] state-of-the-art Bayesian background generation and foreground detection experiments have witnessed up to 20 times performance boost.

References

[1] S.A.Velastin, P.Remagnino. Intelligent distributed video surveillance system. The Institution of Electrical Engineers, New York, pp. 1-2. 2006.

[2] E. H. Aldelson, J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal Optical Society of America*, vol. 2, no. 2, pp. 284-299, 1985.

[3] H. H. Baker, R. C. Bolles. Generalizing epipolar plane image analysis on the spatio-temporal surface. *IJCV*, vol. 3, no. 1, pp. 33-49, 1989.

[4] Y. Li , C. K. Tang, H. Y. Shum. Efficient dense depth estimation from dense multi-perspective panoramas. *IEEE ICCV 2001*, Canada, vol. 1, no. 1, pp. 119-126, 2001.

[5] G. Kuhne, S. Richter, M. Beier. Motion-based segmentation and contour based classification of video objects. *The 9th ACM international conference*, USA, vol. 9, no. 1, pp. 41-50, 2001.

[6] C. W. Ngo, T. C. Pong, H. J. Zhang. Motion analysis and segmentation through spatio-temporal slice processing. *IEEE Trans.IP*, vol. 12, no. 3, pp. 341-355, 2003.

[7] Hirahara, Z. Chenfhua, K. Ikeuchi. Detection of street-parking vehicles from panoramic street image. *IEEE proceedings of ITS*, vol. 2, no. 1, pp. 993-998, 2003.

[8] R. Mandelbaum, G. Salgian, H. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. *IEEE ICCV 1999*, Greece, vol. 1, no. 1, pp. 544-551, 1999.

[9] H. Kawasaki, M. Murao, K. Ikeuchi, M. Sakauchi. Enhanced navigation systems with real images and real-time information. *The 8th World Congress on Intelligent Transport Systems*, Sweden, pp. 221-228, 2001.

[10] A. Ravacha., P. Peleg. A unified approach for motion analysis and view synthesis. *The 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, Greece, vol. 1, no. 1, pp. 717-724, 2004.

[11] M. blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri. Actions as space-time shapes. *IEEE ICCV 2005*, China, vol. 2, no. 1, pp. 1395-1402, 2005.

[12] A. Yilmaz, M. Shah. Actions as objects: a novel action representation. *IEEE Computer Society Conference on CVPR 2005*, USA, vol. 1, no. 1, pp.984-989, 2005.

[13] V. Moolani, R. Balasubramanian, L. Shen, A. Tandon. Shape analysis and spatio-temporal tracking of mesoscale eddies in miami isopycnic coordinate ocean model. *International Symposium on 3D Data Processing Visualization and Transmission*, USA, vol. 1, no. 1, pp. 663-670, 2006.

[14] L. Orelick, M. Alun, E. Haron. Shape representation and classification using the poisson equation. *IEEE trans. PAMI*, vol. 28, no. 12, pp. 1991-2005, 2006.

[15] A. F. Bobick, J. W. Davis. The recognition of human movement using temporal templates. *IEEE trans. PAMI*, vol. 23, no. 3, pp. 257-267, 2001.

[16] D. Weinland, R. Ronfard, E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249-257, 2006.

[17] T. Ogata, J. K. Tan, S. Ishikawa. High-speed human motion recognition based on a motion history image and an eigenspace. *Transactions on Information and System*, vol. E89-D, no. 1, pp. 281-289, 2006.

[18] C. Ware, G. Franck. Evaluating stereo and motion cues for visualizing information nets in three dimensions. *ACM Transactions on Graphics*, vol. 15, no. 4, pp. 121-140, 2006.

[19] M. Xiao, C. Han, L. Zhang. Moving shadow detection and removal for traffic sequences. *International Journal of Automation and Computing*. vol. 4, no. 1, pp. 38-46, 2007.

[20] K. Yamamoto, R. Oi. Color Correction for multi-view video using energy minimization of view network. *International Journal of Automation and Computing*. vol. 5, no. 3, pp. 234-245, 2008.

[21] K. P. H. Berthold, G. R. Brian. Determining Optical Flow. *Artificial Intelligence*, vol. 59, no. 1-2, pp. 81-87, 1993.

[22] K. Michael, W. Andrew, T. Demetri. Snakes: Active contour models. *IJCV*, vol. 1, no. 4, pp. 321-331, 1988.

[23] K. Fukunaga, L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans.IT*, vol. 21, no. 1, pp. 32-40, 1975.

[24] R. O. Duda, P. E. Hart. Pattern classification and scene analysis. Wiley-Interscience, New York, pp.135-137, 2000.

[25] D. Comaniciu, P. Meer. Mean Shift: A robist approach toward feature space analysis. *IEEE Trans. PAMI*, vol. 25, no. 5, 2002.

[26] D. Comaniciu. Nonparametric robust methods for computer vision. PhD thesis, ECE Department, Rutgers University, July 2001.

[27] D. A. Forsyth, J. Ponce. Computer vision: a modern approach. Prentice Hall, pp. 309-313, 2003.

[28] J. Canny. A computational approach to edge detection. *IEEE Trans.PAMI*, vol. 8, no. 6, pp. 679-698, 1986.

[29] F. Porikli. Constant time O(1) bilateral filtering. *IEEE Computer Society Conference on CVPR 2008*, USA, vol. 1, no. 1, pp. 1-8, 2008.

[30] M. Hussein, F. Porikli, P. Meer. Learning on lie Group for invariant detection and tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, USA, vol. 1, no. 1, pp. 2-16, 2008.

Jing Wang received the B. Sc. from Xidian University, Xi'an, China, in 2006. He then joined in Beijing Zhong Ke Fan Hua Measurement & Control Technology Co., Ltd. (also known as Pansino Ltd.) as one of development engineers working on machine vision system for 2 years. He is currently a research student at the Informatics research group of University of Huddersfield, UK.

His research interests include image processing, machine vision and intelligent computer vision system.

E-mail: j.wang2@hud.ac.uk

Dr. Zhijie Xu is a Reader and the head of the Computer Graphics and Image Processing Research Group within the School of Computing and Engineering at the University of Huddersfield. He obtained his PhD in Virtual Manufacturing in 2000 at the University of Derby in United Kingdom. His research has mainly been focused on the

areas of real-time graphics and vision systems, Virtual Reality (VR), manufacturing simulations, and Web-based e-Technologies.

He has published over 50 papers in the relevant fields and served as editorial board, guest editor, and appointed reviewer for various international journals in the engineering and computing domains. He is a charted electronic engineer and a member of the IEEE, IET/IEE, British Computer Society (BCS), and UK Higher Education Academy (HEA)

E-mail: z.xu@hud.ac.uk