AN INNOVATIVE VOLUME BASED VIDEO FEATURE EXTRACTION TECHNIQUE

Jing Wang[1], Zhijie Xu[1], Jonathan Pickering[1] and Rakesh Mishra[1]
[1]University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK

ABSTRACT

Based on the video frames, a spatial-temporal volume data structure represents more flexible processing methods than traditional 2D sequential images approach in computer vision. This paper describes the data structure of the spatial-temporal volume and the feature volume coming from the original video data. A compressed volume structure called a feature video is presented showing how a feature volume can be build based on the sequential frames. The existent 3D volume processing methods such as slice processing, 3D filters and CFD methods are also introduced in this paper. As a practical application, human gait analysis based on volume slice processing is described. This includes the video capture, volume organized and the slice feature extraction. The result of experimental data shows the different features between different behaviors of gait.

Keywords video processing, video volume, feature extraction

1 INTRODUCTION

Using processing methods, the image sequences such as video are mainly processed frame by frame. This is a time consuming procedure. Recently, a volume based video data structure that could extract the feature from video sequence has been introduced. This 3D data structure contents the spatial and temporal information coming from the original video data. The volume structure is called as spatial-temporal volume (STV). A visualization of a STV without processing is showed in Fig 1. One of the major advantages of this representation is that by analyzing feature structures in this volume, we may reason about much longer-term dynamics. Also, by jointly providing spatial and temporal continuity, the complexity of feature correspondence is significantly reduced. A further advantage is that occlusion events are made much easier to detect, as they are represented explicitly in this volume as truncated paths [1, 2].

The paper reports work on the flow signature abstraction and feature identification based on the video volume approach which breaks through the traditional video processing techniques. This approach brings a new point of view about the image flow processing. The research should produce more stabile and effective result than traditional methods.

This paper is organized as follows: Section 2 presents a brief review about the development of the video volume data processing. In Section 3, a developed video volume called feature volume is described. It gives a particular description of the structure and the implement of this special volume data structure. Section 4 presents an example of extracting the video features from a human gait in order to shows the typical feature volume processing method. Section 5 is a summary and presentation of future research.

2 LITERATURE REVIEW

Most video processing methods focus on the spatial information and the temporal relationship between two consecutive frames. Generally speaking, these approaches could be summarized as online and offline methods. Online techniques search abutting frames for some useful information concerning the moving object such as optical flow or feature tracking. Offline processing methods often compute the entire pixel database from the video, for example, rebuilding the background from a series of frames.

The approaches such as difference, or optical flow, share the same characteristic of typically determining motion based on two frames in the image sequence. Motion estimation using frame differencing is highly sensitive to noise and the results have a high false positive rate which is hard to surpress [3]. Feature tracking follows a sparse set of salient image points over many frames, whereas optical flow estimates a dense motion field from one frame to the next [4]. The disadvantage of these

approaches is the motion estimates that are both short-range and unstable feature description.

The volume data structure was introduced to emphasize the temporal continuity and change from a video data. The use of spatial-temporal volumes were first pioneered in 1985 by Aldelson and Bergen [5], who used motion models based on energy and impulse response to filters. There are many existed methods for analyzing this 3D data. One way to analyze the STV is to consider it as being formed by a stack of two-dimensional temporal slices. Three basic volume slices have been built. As showed in Fig2.

1. XY shows the traditional frame from the video. It presents the visual information and the feature distributing at a particularly time.
2. XT slice means a spatial-temporal slice and emphasizes the changes happened on the horizon. This slice usually shows some bars and ripples. The special texture contains some important information comes from original video data according to different applications.
3. YT slice shows the same meaning but puts the basic information gathered from the vertical direction.

These slices have been studied in for a variety of problem domains: to infer feature depth information [6], generating dense displacement fields [7], camera work analysis [8], motion categorization [9], the detection of parked vehicles [10], ego-motion estimation [11], for use in advanced navigation systems [12] and view synthesis [13].

Another way to analyze the STV is to process the 3D data directly. In the approach the spatial-temporal volume for scene segmentation has been used to process the structures found in the entire volume, rather than by analyzing distinct slices. A variety of techniques for doing so have been studied. These include: spatial-temporal manifolds [14], the 3D structure tensor [15], mean shift analysis [16], Fourier analysis [17] and deformable shape models [10]. Very recently, level set evolution equations have been successfully used for spatial-temporal segmentation [18]

3 FEATURE VOLUME STRUCTURE

Basically, all the STV data is the 3D data. As showed in Fig 1, original STV data comes from the video frames one by one according to the temporal order. This reorganization keeps every pixel and transforms them into the 3D space which may be realized as a kind of *voxels* space. That means features in this volume are not extracted. All the data had to be processed after the 3D volume are built. This logic could introduce some problems. One significant problem is that the STV data space could be too big to deal with. The size of this volume has relationship with the frame size and the time durance of the video. For example, a 320*240 pixels image with 150 slices in the volume should consume 33MB memory space in the computer. This is only the 5 seconds video frames if the fps=30. Processing this huge volume is also a time consuming procedure.

A new feature volume structure is introduced in order to solve this problem. This feature volume has two important parts, the pre-processing frame data and the compressed volume structure. The new feature volume only keeps useful feature in each frame, which means before the 3D volume is organized, some traditional image processing methods has been used in each frame for the rough feature coming from the images. Useful image processing methods such as optical flow or partner recognition approaches could extract the rough feature from the frames. This method removes the useless pixels and abstracts the useful feature dependant on the application. It also reduces the data volume and more important, building the 3D feature volume and the pre-processing of the video frame is a parallel processing if the *producer-consumer structure* [19] is used.

Notice this feature volume has a low level of the entropy. The compressing technology could give an approximate result. It takes advantage of an opportunity to organize the feature slice by the time order into a video and use the video compression technology for the feature volume. That means some popular compression method and file structures can be used in the volume compression applications. The meaningless area in the video could be compressed automatically based on different file format and normative code such as MPEG or DVIX. Compared with the example above, the 320*240 with 150 slices volume costs only 130KB if it is an AVI file and the DVIX code.

4 ONE APPLICATION OF THE FEATURE VOLUME

As a basic data for the volume processing, the feature volume could be used as input for the volume feature extraction. In order to explain the details about the typical processing steps, a human gait analysis based on the volume slice is introduced as an example. The target of this application is extracting the features of the gait. Some basic features such as walking speed and cycle need to be extracted.

The slice is one approach in the volume processing method that has been mentioned in Section 2. Compared with the slice coming from original STV, the human gait analysis is based on the feature volume. Further processing is mainly based on the feature slice from the XT-planes. The following parts of the section present each step according to this application.

Capturing the video footage
The video is captured in a sunny outdoor environment. The content of this video is a walker crossing a road. Fig 3 shows a snapshot coming from this capture. We captured three different videos. That is walking from left to right, walking from right to left and running from left to right. The frame size is 720*480. The frame rate is 25 frames per second. In this approach, the camera is fixed, the walker walks at mostly a constant speed, the direction of walk is roughly lateral relative to the camera, and no obstacles carried by the walker are present.

Building the feature video
Based on the method mentioned in Section 3, this feature video is taken from processed frames. Some image processing methods such as image edge strengthen and noise filter has been used in this application. Simple threshold remove unnecessary pixel and change them into 0. The pre-processing pixels are the basic data of the feature volume and compressed as a feature video.

Building the XT feature slice
One prototype based on the Open CV and VC++ has been build. Based on the feature video extracted from the original video, the feature slices could be composed directly from the existent volume cube. That means the video volume feature data saved as a compressed feature video should be decompressed and provided the information required to build the feature slice. Fig 4 shows the basic logic of the prototype for building a feature slice. The time consuming of this operation is based on the duration of the feature video, the frame rate and the size of the frame.

The XT-slice indicates that the head undergoes pure translation during normal walking. Fig 5a shows an XT-slice obtained near the walker's head; An XT-slice of the volume near the walker's ankle reveals a unique braided signature for walking patterns. Fig 5b shows the walker's legs crossing over one another as the walker walks from left to the right. These braided patterns are generated by all human walkers.

The XT slice near the human head explains the main track and direction of the human. This texture reflects the head movement during a special time portion. It shows the main body movement because the experience that the head always keep the same speed and direction with the body. The straightness shows whether the person moves at a symmetrical speed. The size of the angle shows the speed of the walker. As an especial example, the static object in that XT slice shows as some vertical bars. That means the angle of the line with the horizontal direction reflect the speed of the walker. The straightness shows the dynamic process of the transform of the speed.

XT feature slice processing
In order to get a clear line from the complicated bars, the vertical bars need to be reduced automatically. One effective method is judged the vertical bars as some kind of background. This idea is based on the traditional methods that the static object in the images should be realized as the background and the static object always shows as the vertical bars in the XT slices. This background is easier to build in the XT slice because this slice contains the temporal information natively. That means we could build a background with some 2D image processing methods rather than frames difference methods.

One appropriate method is a statistical median along with the temporal direction. Usually, this method needs to ransack the whole video to get only one median value. Because the slice has been built when the feature volume is established, this XT slice could be processed as an image with a traditional

median filter. The size of this filter is the *h*1 (pixels)*. The *h* denotes the number of frames. Based on the same theory, the background of the XT slice near the ankle could also be processed with the same median filter. Fig 6a and Fig 6b shows the background built by the median filter. Fig 6c and Fig 6d shows the slices which are remove the background based on the absolute images difference.

Feature analysis and experimental results
The main processing steps of the head slices are based on the line fitting. The result of this fitted line is shows in Fig 7. This line fitting method is based on the edge detection. This approach locates a straight edge in a rectangular search area. The method locates the intersection points between a set of parallel search lines and the edge of an object. It determines the intersection points based on their contrast and slope and calculates a best-fit line based on these points.

The edge detection method is a simple detector of the edge based on the threshold. The method uses the pixel value at any point along the pixel profile to define the edge strength at that point [21]. In order to find and fit the head slice whit a straight line, the fitting threshold is set to 70.

The main processing steps of the ankle slices and the results are showed in Fig 8. In this application, removing small particles from the ankle slice is based on low pass and high pass filters [21]. After removing the small particles, the ankle slice shows a series of footprints from a top view. This processed slice presents a spatial X-component and temporal track.

This cycle is easy to get after measuring the coordinates of the objects. Three different kinds of video are measured respectively. Table 1 shows details of these video features. The different speed and different direction of the walker could be easily judged from the characters of these data. The larger angle of the best fit-line shows the higher speed of the walker. The greater straightness shows greater constancy. The spatial cycle shows the span of the two legs and the temporal cycle show the frequency of the alternation of legs.

These feature data could be used in different high-level application. For example, a stick model of a person could be easily built based on these data. Stick models could be used to analyze the movement of athletes in order to improve their motion. Furthermore, distinguishing the normal and deviant behavior in surveillance system could also benefit from this kind of stick models.

5 FUTURE WORK AND SUMMARY

Research on the processing methods of the feature volume is needed. Existed 3D STV methods only focused on the traditional image data, which means 2D feature slices and 3D filters are also a development of the original image processing methods. Feature slices mapping the 3D spatial and temporal volume into a 2D environment and manages the data as traditional image data. On the other hand, 3D filters are all developed from existed 2D filters.

Beyond the image processing area, many other researchers pay more attention on the 3D feature data. For example, Computational Fluid Dynamics (CDF) research area shows many practical approaches for the handling of 3D fluid features. These approaches could transplant into the computer vision area for the 3D video features. The only difference is that CFD data source comes from sensors while feature volume data source comes from different image processing systems. After mapping the feature volume in a CFD 3D matrix, we could apply some CFD models on the video feature extraction. In order to serve data for CFD models, the feature volume must be a 3D vector field. An ideal vector field in the image processing area is optical flow. Pushing the 2D optical flow fields into the feature volume builds a 3D vector field. The third dimension of optical flow is time value.

For example, given a velocity vector $\vec{V} = (u, v, w)^T$ equal to the flow velocity we may write the Navier-Stokes equations. These equations are based on the principle of conservation of mass, momentum and energy and are presented in continuity equation (1), momentum equation (2) and energy equation (3). These CFD models are derived from Newton`s second law and the first law of thermodynamics. [22]

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \qquad (1)$$

$$\frac{D\vec{V}}{Dt} = -\frac{1}{\rho}\nabla p + \nu\nabla^2\vec{V} + \vec{f} \quad (2)$$

$$\rho c_p \frac{DT}{Dt} = -\left(\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} + \frac{\partial q_z}{\partial z}\right) + \dot{q}_h \quad (3)$$

These conservation equations contain the surface forces, viscous stress and energy information based on the velocity vector and the Navier-Stokes equations. Mapping these theoretic models into feature video with optical flow vector field may change the physical meaning in the real-world. The force, pressure, viscous stress and energy values and their rate of changes could introduce correlative values which characterize the significant features in the optical flow vector field. Different distributions of the flow vectors show different feature value in the CFD models. That could be used in the video segmentation approach if the classifications of the optical flow vectors are needed.

The volume based video feature extraction introduces an innovative approach for the video analysis. This approach pays more attention on the temporal changes and the relationship of the moving objects between different frames. A significant feature could be extract directly from the volume based on 3D data processing methods rather than traversal in the whole video by 2D image processing methods. The effective feature volume contains different features in different applications. The useful feature could save as a compressed file in the 3D feature video format. The deep processing methods based on this feature volume usually bring more feature than frame processing because the additional temporal information and so many reliable 3D signal processing methods.

REFERENCES

[1] CONRAD J. and RISTIVOJEVIC M. (2004), *Joint Space-Time Image Sequence Segmentation: Object Tunnels and Occlusion Volumes*. International Conference on Acoustics, Speech and Signal Processing, pp. 9-12.

[2] SWAMINATHAN R., KANG S.B., CRIMINISI A. and SZELISKI R. (2002), *On the Motion and Appearance of Specularities in Image Sequences.* European Conference in Computer Vision.

[3] COLLINS T. (2004), *Analyzing Video Sequences Using the Spatio-Temporal Volume.* MSc Informatics Research Review.

[4] SAND P. and TELLER S. (2006), *Particle Video: Long-Range Motion Estimation Using Point Trajectories*, CVPR, pp. 2195-2202.

[5] ALDELSON E. and BERGEN J.R. (1985), *Spatiotemporal Energy Models for The Perception of Motion*. Journal Optical Society of America, Vol.2, pp. 284-299.

[6] Baker H.H. and Bolles R. C. (1988), *Generalizing Epipolar Plane Image Analysis on the Spatio-Temporal Surface*. n Proceedings of the DARPA Image Understanding Workshop, pp.33-48.

[7] LI Y., TANG C.K. and SHUM H.Y. (2001), *Efficient Dense Depth Estimation from Dense Multi-Perspective Panoramas*. International Conference on Computer Vision, pp. 119-126.

[8] KUHNE G., RICHTER S. and BEIER M. (2001), *Motion-Based Segmentation and Contour based Classification of Video Objects*. The 9th ACM international conference.

[9] NGO C.W., PONG T.C. and ZHANG H.J. (2003), *Motion Analysis and Segmentation through Spatio-Temporal Slice Processing*. IEEE Transactions on Image Processing, pp. 341-355.

[10] HIRAHARA K., CHENGHUA Z. and IKEUCHI K. (2003), *Panoramic-View and Epipolar-Plane Image Understandings for Street-Parking Vehicle Detection*. ITS Symposium.

[11] ONO S., KAWASAKI H., HIRAHARA K. and KAGESAWA, M. (2004), *Ego-Motion Estimation for Efficient City Modeling Using Epipolar Plane Range Image Analysis*, in TSWC2003.

[12] KAWASAKI H., MURAO M., IKEUCHI K. and SAKAUCHI M. (2004), *Enhanced Navigation Systems with Real Images and Real-Time Information*. International Journal of Computer Vision, vol. 58, No.3, pp.237-247.

[13] RAV-ACHA A. and PELEG P. (2004), a Unified Approach for Motion Analysis and View Synthesis. 2nd International Symposium on 3D Data Processing, Visualization, and Transmission,

pp.717-724.

[14] BAKER H.H. and GARVEY T.D. (1991), Motion Tracking on the Spatiotemporal Surface. The IEEE Workshop of Visual Motion, pp.340-345.

[15] NGO C.W., PONG T.C. and ZHANG H.J. (2000), *Motion Characterization by Temporal Slice Analysis*. IEEE Conference on Computer Vision and Pattern Recognition.

[16] DEMENTHON D. and MEGRET R. (2002), *Spatio-Temporal Segmentation of Video by Hierarchical Mean Shift Analysis*. Statistical Methods in Video Processing Workshop, pp.800-810.

[17] OHARA Y., SAGAW R., ECHIGO T. and YAGI Y. (2004), *Gait Volume: Spatio-Temporal Analysis of Walking*. The fifth Workshop on Omni directional Vision, Camera Networks and Non-classical cameras, pp.79-90.

[18] MITICHE A., FEGHALI R. and MANSOURI, A.(2003), *Motion Tracking as Spatio-Temporal Motion Boundary Detection*, Journal of Robotics and Autonomous Systems, Vol. 43, pp.39-50.

[19] GREGORY R. A. (2000), *Foundations of Multithreaded, Parallel, and Distributed Programming.* ISBN. 0201357526, pp.27-69.

[20] Open CV Wiki-pages, http://opencvlibrary.sourceforge.net

[21] NI Developer Zone, *Line profile,* http://zone.ni.com/devzone/cda/epd/p/id/5560

[22] TUNCER C., JIAN P.S., FASSI K. and ERIC L. (2005), *Computational Fluid Dynamics for Engineers.* HoriZones Publishing, Springer, ISBN. 3540244514, pp.42-46.

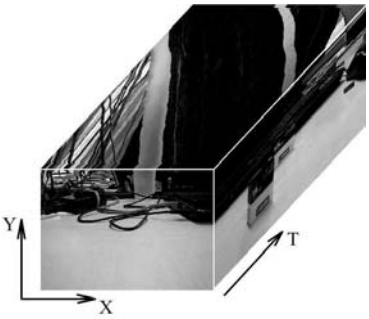

Figure1: 3D Video volume structure

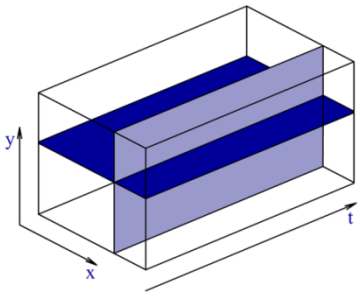

Figure 2: STV slices

Figure 3: Footage of the human gait (one frame from a video)

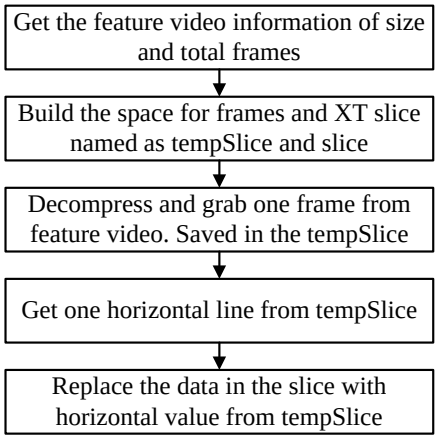

Figure 4: Flowchart of building a feature slice

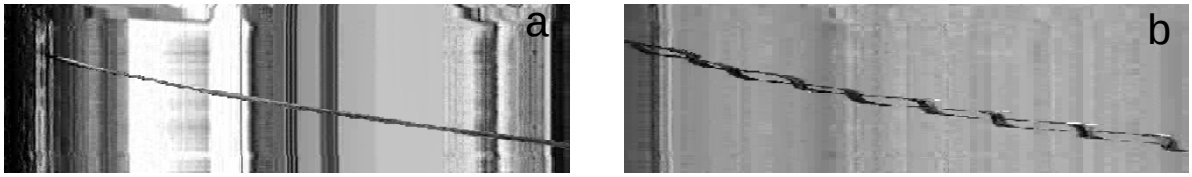

Figure 5: Two feature slices coming from gait video

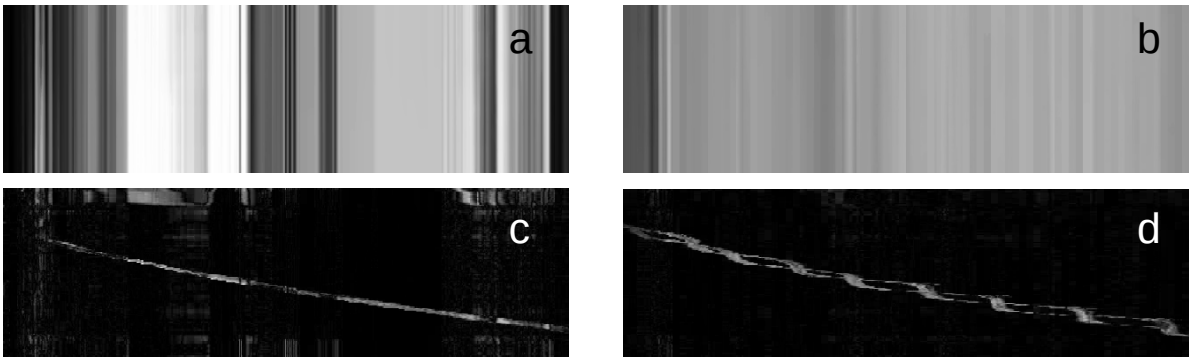

Figure 6: Processing of the gait slices

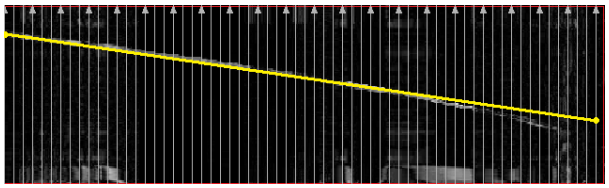

Figure 7: The fitting result

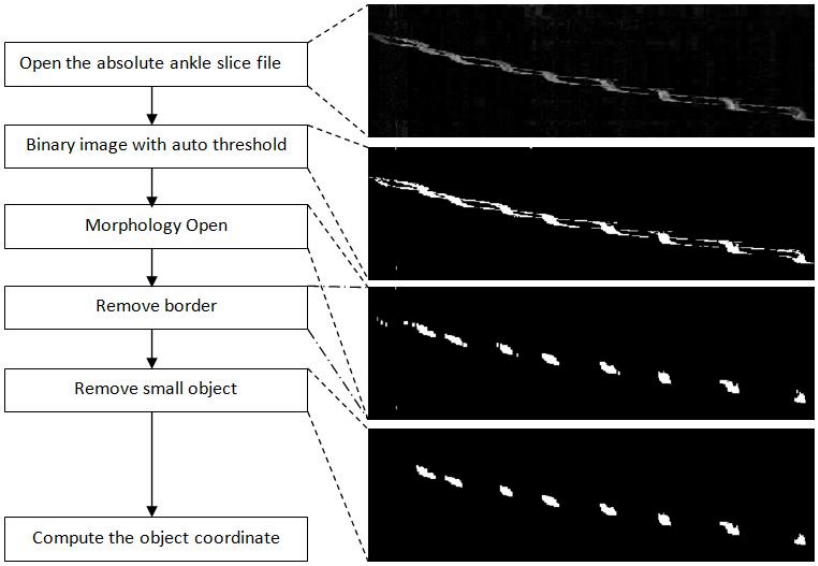

Figure 8: Processing steps and result of the human ankle slice

	Walk from left to right	Walk from right to left	Run from left to right
Angle (degree)	12.15	-10.36	21.47
Straightness (%)	79.66	81.51	61.35
Average spatial cycle (pixels)	70	-66	112
Average temporal cycle (pixels)	13	-14	16

Table 1: Result of three different videos.