Volume-based Video analysis using 3D Segmentation Techniques

Jing Wang, Zhijie Xu, Jonathan Pickering
Department of Informatics, School of Computing and Engineering
University of Huddersfield
Queensgate, Huddersfield HD1 3DH, United Kingdom
j.wang2@hud.ac.uk, z.xu@hud.ac.uk, j.h.pickering@hud.ac.uk.

Abstract— **Video content understanding for surveillance and security applications such as smart CCTV cameras has become a research hotspot in the last decade. This paper presents applications of volume construction and volume-based cluster segmentation approaches to video event detection. It starts with a description of the translation between original video frames and 3D volume structures denoted by spatial and temporal features. It then highlights the volume array structure, a so called "pre-suspicion" mechanism. The focus of the work is on devising an effective and efficient voxel-based segmentation technique suitable to the volumetric nature of video events through deploying innovative 3D clustering methods. It is supported by the design and experiment on the 3D data compression techniques for accelerating the pre-processing of the original video data. An evaluation on the performance of the developed methods is presented at the end.**

Keywords - spatio-temporal volume; video processing; volume feature extraction; segmentation

I. INTRODUCTION

Inherited from image processing techniques, traditional video event detection approaches puts 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 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 some "changes" occurred in a Euclidean space over a period of time elapsed. Both the recorded spatial and temporal signals can be either 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 array 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 smallest element inside of the STV "box" is called a voxel, which holds the pixel and the time information together.

The STV data structure transforms video event detection approaches from a FBF mechanism to one of a 3D incorporate shape analyses. Useful events can be extracted directly from the volume by deploying appropriate feature matching approaches. This process mainly relies on image segmentation processes, which had been well developed by image processing community. As shown in Fig.2, a "waving hand" event can be extracted to form a STV model. It shows the feature segmentation operation that highlights the contour of the non-rigid human body changes and also denotes the original STV with two labels: interesting features and background.

If different video event can be abstracted and modeled as 3D template shapes, then the corresponding event detection tasks can be reduced into the jobs of recognizing the 3D shapes in any 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 which denote certain features.

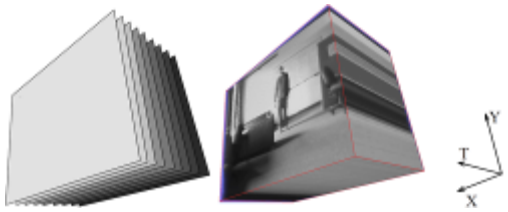

Figure 1. STV strcture

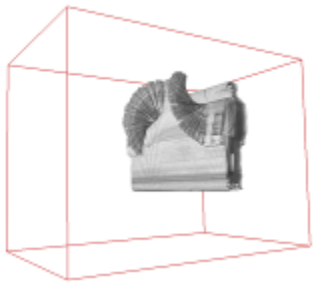

Figure 2. STV model of a waving hand

This project develops two technologies. The first one focuses on the volume event feature extraction, which forms the core of this paper. The second one solves 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 modelling 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 some experimental results. The result is further analyzed and discussed in Section 5. Section 6 covers the conclusions and future works.

II. Literature Review

The volume data structure mentioned earlier emphasizes the temporal continuity in an input stream of a video data. The use of spatial-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 benchmarking or thresholding 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, 12, 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].

III. STV Construction and STV Array Structure

STV is a 3D volume data structure, which is widely used in medical visualizations, such as 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.

A. 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 $F_i(1,2,\ldots,n)$ 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 $\mathbf{p} = (p_j, p_k)$ where *j* and *k* denote the coordinate values of the pixel in the 2D image plane. The function $I = I(\mathrm{p})$ 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 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.

B. 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.

C. 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 looking 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 [19] 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.

Appropriate compressing technologies can further reduce the memory footprint. The latter part of the devised process applies an 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 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 DVIX code.

Figure 3. Pre-suspicion STV array mechanism

IV. Voxel-based Feature 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 stop 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 [20]. 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.

Since the clustering methods in general intend to sort the studied elements by the pre-defined spectrums, in terms of volume studies, voxels sharing similar 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 form 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 [21] as a nonparametric method to estimate a Probability Density Function (PDF) using the so-called Parzen window density estimator [22]. Using a similar notation as explained in [23], the MS technique can be described as follows: given n data points $\mathbf{x}_i$, i=1,...,n in the d-dimensional feature space, the multivariate kernel density estimator with kernel $K_H(\mathbf{x})$ computed at the point $\mathbf{x}$ is given by

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

Comaniciu et al. [23] discussed that for the mean shift vector $m_{h,G}(\mathbf{x})$ the following condition holds

$$m_{h,G}(\mathbf{x}) = \frac{h^2 c}{2} \frac{\nabla f_{h,K}(\mathbf{x})}{f_{h,G}(\mathbf{x})}, \quad (4)$$

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. These profiles are related by the condition

$$g(\mathbf{x}) = -k'(\mathbf{x}), \quad (5)$$

where k' is the derivative of the profile k. (4) 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 [24]. 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 [22] 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.

V. Experiments Result

To assess the devised STV feature model and the corresponding segmentation by clustering approach, a set of experiments have been designed and carried out. The software tools and APIs 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.

A. 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 was as only interested in moving objects in the video and then composes STV shapes by assembling moving contours. As mentioned in Section 3, these steps must be finished 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 the 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 [25] and was calculated by capturing 100 frames with 100ms alternation.

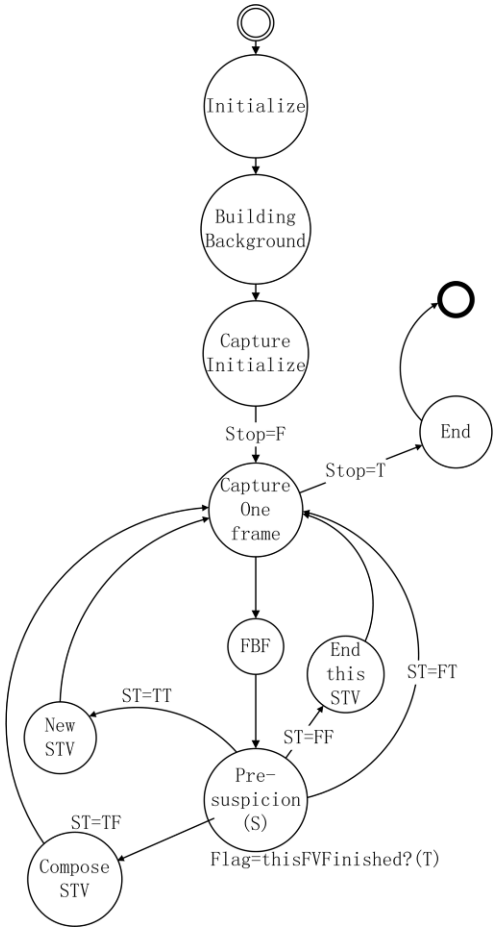

Figure 4. Pre-suspicion algorithm

Following the "capture one frame", a FBF image processing was used to find the canny edge [26] for the contour in the absolute differences image. The high and low-threshold is 70% and 30% of the maximum pixel value. The size of the Gaussian smoothing filter was 9 by 9, the result of which is shown in the Fig.5.

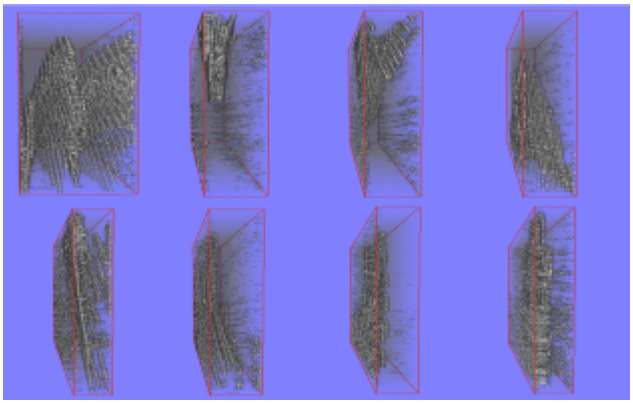

Figure 5. Result of STV array

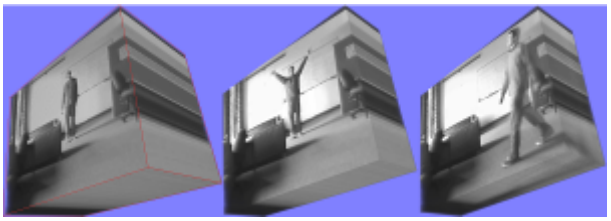

Figure 6. Original STV contains different events

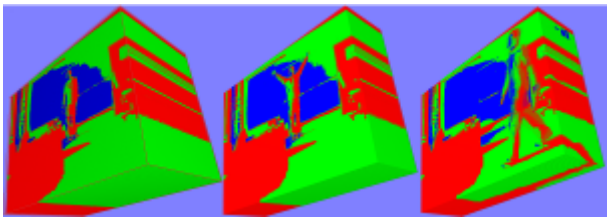

Figure 7. 3D K-Mean Segmentation result

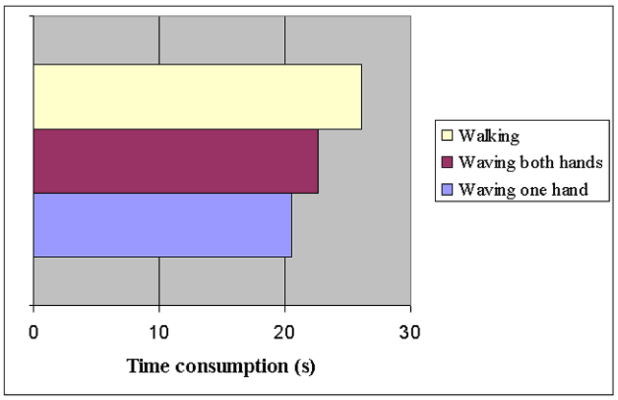

Figure 8. Time consumption of voxel-based K-Mean segmentation

Another important features in this algorithm is the logic flags which control the state transition and are marked as "Pre-suspicion?" and "Is this STV finished?" in Fig.4. After evaluating whether the current frame contains moving contours in the "Pre-suspicion" phase, different process combinations of these flags will lead to different transition directions.

B. Voxel-based Segmentation

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. As shown in Fig.6, They are "Waving one hand", "Waving both hands" and "Walking" events.

The K-Mean method adopted in this experiment is based on the intensity of the gray-level for each voxel. The approach is an upgrade from the 2D-based pixel operation since the only difference is the extra dimension introduced by the voxel which can be readily handled by the vector expression of many classic clustering algorithms. The result is illustrated in Fig.7. The time consumption of the operation 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 consumption is substantial. Some anticipate solutions for alleviating this problem will be discussed in the final section of this paper.

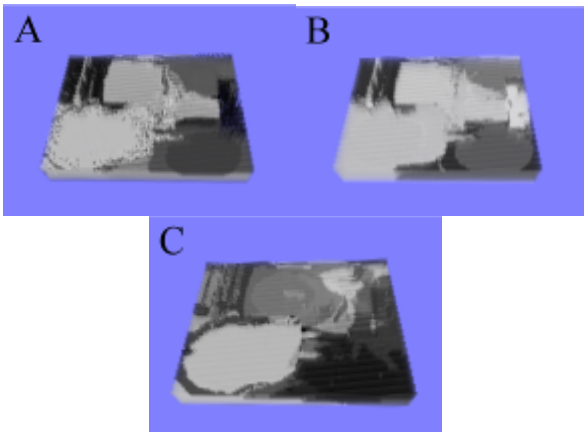

Figure 9 Voxel-based MS operation

The Mean Shift (MS) clustering technique was also experiment 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 [23]. The result is shown in Fig.9.A, 9.B and 9.C (with both the H_r and H_s set at 32) representing the waving one hand, waning both hands and walking events, respectively. The overall time consumption is over 300 seconds.

VI. 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 report 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. With various 2D data processing techniques such as the K-Means and MS, various clustering segmentation approaches have being successfully transformed into 3D volume space. The key task in the future is to devise template mapping techniques which can be applied to the 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 the more complex operation, such as the Mean-Shift, which contains many iterative steps, the run time of the algorithmic becomes 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 [27]. It is understood in this research that most STV processing techniques handle each voxel the same arithmetic operation, which can be realized in programmable GPU streams one of the parallel data processing mode – SIMD (Single Instruction Multiple Data). The acceleration factor has been proven in many early studies. For example, comparing to the CPU-dominant approach, the Meer's [28] state-of-the-art Bayesian background generation and foreground detection experiments has witnessed a 20X performance boost.

REFERENCES

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

[2] E.Aldelson and J.R.Bergen, "Spatiotemporal energy models for the perception of motion," Journal Optical Society of America, Vol.2, 1985, pp. 284-299.

[3] H.H.Baker and R.C.Bolles, "Generalizing epipolar plane image analysis on the spatio-temporal surface," n Proceedings of the DARPA Image Understanding Workshop, 1988, pp. 33-48. 1988.

[4] Y.Li , C.K.Tang and H.Y.Shum, "Efficient dense depth estimation from dense multi-perspective panoramas," ICCV, VOl.1, 2001, pp.119-126.

[5] G.Kuhne, S.Richter and M.Beier, "Motion-based segmentation and contour based classification of video objects," The 9th ACM international conference, 2001.

[6] C.W.Ngo, T.C.Pong and H.J.Zhang, "Motion analysis and segmentation through spatio-temporal slice processing," IEEE Trans.IP, Vol.12, 2003, pp. 341-355.

[7] Hirahara, Z.Chenfhua. and K.Ikeuchi, "Panoramic-view and epipolar-plane image understandings for street-parking vehicle detection," ITS Symposium, 2003.

[8] S.Ono, H.Kawasaki, K.Hirahara and M.Kahesawa, "Ego-motion estimation for efficient city modeling using epipolar plane range image analysis," in TSWC 2003, 2004.

[9] H.Kawasaki, M.Murao, K.Ikeuchi and M.Sakauchi, "Enhanced navigation systems with real images and real-time information," IJCV, vol.58, 2004, pp. 237-247.

[10] A.Rav-acha. and P.Peleg, "A unified approach for motion analysis and view synthesis," 2nd International Symposium on 3D Data Processing, Visualization, and Transmission, 2004, pp. 717-724.

[11] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," ICCV, 2005.

[12] A. Yilmaz and M. Shah, "Actions as objects: a novel action representation," CVPR, 2005.

[13] Gorelick and M.Blank, "Actions as space-time shapes," IEEE Trans.PAMI, Vol. 29, 2007, pp. 2247-2253.

[14] L.orelick, M.alun and E.haron, "Shape representation and classification using the poisson equation," IEEE trans.PAMI, Vol. 28, 2006, pp. 1991-2005.

[15] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," PAMI, Vol.23, Issue 3, 2001.

[16] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," Computer Vision and Image Understanding, Vol. 104, Issue 2, 2006.

[17] "Open source computer vision library reference manual," , 2002, pp. 35-36.

[18] C.Ware and G.Franck, "Evaluating stereo and motion cues for visualizing information nets in three dimensions," ACM Transactions on Graphics, Vol. 15, Apr 2006, pp. 121-140.

[19] K.P.H.Berthold and G.R.Brian, "Determining Optical Flow," Artificial Intelligence, Vol. 17, 1981, pp. 185-203.

[20] K.Michael, W.Andrew and T.Demetri, "Snakes: Active contour models," IEEE Trans.IJCV, 1988, pp. 321-331.

[21] K.Fukunaga and L.D.Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Trans.IT, Vol. 21, 1975, pp. 32-40.

[22] R.O. Duda and P.E.Hart, "Pattern classification and scene analysis," Wiley-Interscience, New York, 2nd Edit., 2000.

[23] D.Comaniciu and P.Meer, "Mean Shift: A robist approach toward feature space analysis," IEEE Trans. PAMI, Vol. 25, May 2002, Issue 5.

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

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

[26] J.Canny, "A computational approach to edge detection," IEEE Trans.PAMI, Vol.8, 1986, pp. 679-714.

[27] F.Porikli, "Constant time O(1) bilateral filtering," CVPR, Jun 2008.

[28] M.Hussein, F.Porikli and P.Meer, "Learning on lie Group for invariant detection and tracking," CVPR, Jun 2008.

.