High Definition Video Segmentation Techniques-A Review

R. S. Sabeenian and S. Lavanya

***Abstract*—Real time segmentation of moving regions in image sequences is a fundamental step in many vision systems including automated visual surveillance human-machine interface and very low bandwidth telecommunications. Background identification is a common feature in many video processing systems. One of the most important back ground identification algorithm is the Gaussian Mixture Model algorithm (GMM).On implementation of the Gaussian mixture model on FPGA results in Reduction of the processing capability of the overall system. Trainable Segmentation is adapted to improve the processing capability. After analyzing and evaluating the performance we conclude with several promising directions for future research.**

***Index terms*—Video segmentation, back ground identification.**

I. Introduction

Background modeling [1], is often used in different applications to model the background and then detect the moving objects in the scene like in video surveillance [2], [3], optical motion capture [4]-[6] and multimedia [7]-[10]. The simplest way to model the background is to acquire a background image which doesn't include any moving object. In some environments, the background isn't available and can always be changed under critical situations like illumination changes, objects being introduced or removed from the scene. To take into account these problems of robustness and adaptation, many background modeling methods have been developed and the most recent surveys can be found in [11], [12]. These background modeling methods can be classified in the following categories: Basic Background modeling [13], [14], Statistical Background Modeling [15], Fuzzy Background Modeling [16] and Background Estimation [17]. Other classifications can be found in term of prediction, recursion, adaptation, or modality [18]. All these Modeling approaches are used in background subtraction context which presents the following steps and issues: background modeling, background initialization, background maintenance, foreground detection, choice of the feature size (pixel, a block or a cluster), choice of the feature type (color features, edge features, stereo features, motion features and texture features). Developing a background subtraction method, all these choices determine the robustness of the method to the critical situations met in video sequence [3] Noise image due to a poor quality image source (NI),Camera Jitter (CJ), Camera automatic adjustments (CA),Time of the day (TD), Light Switch (LS), Bootstrapping (B),Camouflage (C), Foreground Aperture (FA), Moved background objects (MO), Inserted Background Objects (IBO), Multimodal Background (MB), Waking Foreground Object (WFO), Sleeping Foreground Object (SFO) and Shadows (S).Different datasets benchmarks are available to evaluate the robustness of the background subtraction methods against these critical situations which have different spatial and temporal characteristics which must be take into account to obtain a good segmentation. We explain the video segmentation algorithms in section ii and back Ground modeling in section iii and proposed solution to the drawback is given in section IV.

Manuscript received January 14, 2013; revised May 20, 2013.

R. S. Sabeenian is with ECE Department in Sona College of Technology, Salem, Tamil Nadu, India. He is currently heading the research group named Sona SIPRO (SONA Signal and Image PROcessing Research Centre) centre located at the Advanced Research Centre in Sona College of Technology, Salem.

S. Lavanya is with Jayam College of Engineering and Technology, Dharmapuri (e-mail: laavanyaprakash@gmail.com).

II. Video Segmentation Algorithms

Chien et al developed an adaptive background model using so called background registration approach. In their paper, they assume that the longer a pixel remains stationary, the more probable that it belongs to the background. By counting whether a pixel stays approximately in the same value for a predefined period, a new background pixel is registered to the background memories where the old value is discarded. In this way, background plane is updated progressively and the moving object is detected by thresholding the difference between the current frame and the registered background plane. A number of other adaptive background models have also been reported. Although these algorithms produce better modeling towards real world scenarios by background learning process, most of them fail to deal with multi-modal background distribution. A multimodal background distribution is caused by repetitive background object motion, for example, swaying trees, reflections of the lake surface, flickering of the monitor etc. As the pixel, lying in the region where repetitive motion occurs, will generally consists of two or more background colors, the RGB value of that specific pixel changes over time. This would result in false foreground object detection by most adaptive background estimation approach mentioned above. In [19], a background model based on multi-modal pixel distribution is proposed to address the issue. By representing each pixel process using a mixture of Gaussian distributions, repetitive background motions are merged into one of the several background distributions for each pixel. However, as the

algorithm processes video stream pixel wise by updating several Gaussian distributions for each pixel, the calculation burden in parameter updating is unbearable for computers in real time applications. In [19], only a frame rate of 11-13 frames/s is obtained even for small frame size of 160×120 on an SGI O2 workstation. For real time video applications with larger frame size, dedicated hardware architecture seems to be a must. However, as far as the aut hor's knowledge, no such hardware implementation has been reported before. Furthermore, issues emerge with regard to memory bandwidth and storage when it comes to implementation, which is quite common to most video/image processing task. Since the update of background distribution is slow most of the time in a slowly changing scene, the word length of each parameter tends to grow to fulfill the increasing dynamic range. With a reasonable frame size of 352×288 that is used throughout this paper, and under the assumption that 3 Gaussian are used for each pixel process, approximately 6 MB data have to be updated for each frame. This imposes a huge demand for calculation as well as memory bandwidth and size. In paper, dedicated hardware architecture is developed aiming to address all the issues mentioned. With an FPGA platform, simulations can be accomplished in real-time to observe long term effects resulting from fixed point quantization as well as parameter settings. In addition, a controller synthesis tool is developed to reduce the design effort for controller design.

III. Background Modeling Using Mixture of Gaussians

In the context of a traffic surveillance system, Friedman and Russell proposed to model each background pixel using a mixture of three Gaussians corresponding to road, vehicle and shadows. This model is initialized using an EM algorithm. Then, the Gaussians are manually labeled in a heuristic manner as follows: the darkest component is labeled as shadow; in the remaining two components, the one with the largest variance is labeled as vehicle and the other one as road. This remains fixed for all the process giving lack of adaptation to changes over time. For the foreground detection, each pixel is compared with each Gaussian and is classified according to it corresponding Gaussian. The maintenance is made using an incremental EM algorithm for real time consideration. Stauffer and Grimson [19] generalized this idea by modeling the recent history of the color features of each pixel$\{X_1 \ldots\ldots X_t\}$ by a mixture of K Gaussians. We remind below the algorithm.

Principle

First, each pixel is characterized by its intensity in the RGB color space. Then, the probability of observing the current pixel value is considered given by the following formula in the multidimensional case:

$$P(Xt) = \sum \omega_{i,t}\, \eta(X_t, \mu_{i,t}, \sum_{i,t}) \quad (1)$$

where the parameters are K is the number of distributions,$\omega_{i,t}$ is a weight associated to the ith Gaussian at time t with mean $\mu_{i,t}$ and standard deviation$\sum_{i,t}\eta$ is a Gaussian probability density function:

$$\eta(X_t, \mu_{i,t}, \sum) = 1/(2\pi)^{n/2} \left| \sum \right|^{1/2} (e^{-(1/2(xt-\mu)\sum (xt-\mu)} \quad (2)$$

For computational reasons, Stauffer and Grimson [19] assumed that the RGB color components are independent and have the same variances. So, the covariance matrix is of the form:

$$\sum_{i,t} = \sigma_{i,t} I \quad (3)$$

So, each pixel is characterized by a mixture of K Gaussians. Once the background model is defined, the different parameters of the mixture of Gaussians must be initialized. The parameters of the MOG's model are the number of Gaussians K, the weight $\omega_{i,t}$ associated to the ith Gaussian at time t, the mean $\mu_{i,t}$ and the covariance matrix $\Sigma_{i,t}$, K determined the multimodality of the background and by the available memory and computational power. Stauffer and Grimson [19] proposed to set K from 3 to 5.The initialization of the weight, the mean and the covariance matrix is made using an EM algorithm. Stauffer and Grimson [19] used the K-mean algorithm for real time consideration. Once the parameters initialization is made, a first foreground detection can be made and then the parameters are updated. Firstly, Stauffer and Grimson [19] used as criterion the ratio $r_{j=}\omega_j/\sigma_j$ and ordered the K Gaussians following this ratio. This ordering supposes that a background pixel corresponds to a high weight with a weak variance due to the fact that the background is more present than moving objects and that its value is practically constant. The first B Gaussian distributions which exceed certain threshold T are retained for a background distribution:

$$B = \text{argmin}_b (\sum \omega_{i,t} \sum T) \quad (4)$$

IV. Trainable Segmentation

Most segmentation methods are based only on color information of pixels in the image. Humans use much more knowledge than this when doing image segmentation, but implementing this knowledge would cost considerable computation time and would require a huge domain-knowledge database, which is currently not available. In addition to traditional segmentation methods, there are trainable segmentation methods which can model some of this knowledge. Neural Network segmentation relies on processing small areas of an image using an artificial neural network or a set of neural networks. After such processing the decision-making mechanism marks the areas of an image accordingly to the category recognized by the neural network. A type of network designed especially for this is the Kohonen map.

Pulse-Coupled Neural Networks (PCNNs) are neural models proposed by modeling a cat's visual cortex and developed for high-performance biomimetic image processing. In 1989, Eckhorn introduced a neural model to emulate the mechanism of a cat's visual cortex. The Eckhorn model provided a simple and effective tool for studying the visual cortex of small mammals, and was soon recognized as having significant application potential in image processing. In 1994, the Elkhorn model was adapted to be an image

processing algorithm by Johnson, who termed this algorithm Pulse-Coupled Neural Network. Over the past decade, PCNNs have been utilized for a variety of image processing applications, including: image segmentation, feature generation, face extraction, motion detection, region growing, noise reduction, and so on. A PCNN is a two-dimensional neural network. Each neuron in the network corresponds to one pixel in an input image, receiving its corresponding pixel's color information (e.g. intensity) as an external stimulus. Each neuron also connects with its neighboring neurons, receiving local stimuli from them. The external and local stimuli are combined in an internal activation system, which accumulates the stimuli until it exceeds a dynamic threshold, resulting in a pulse output. Through iterative computation, PCNN neurons produce temporal series of pulse outputs. The temporal series of pulse outputs contain information of input images and can be utilized for various image processing applications, such as image segmentation and feature generation. Compared with conventional image processing means, PCNNs have several significant merits, including robustness against noise, independence of geometric variations in input patterns, capability of bridging minor intensity variations in input patterns, etc.

Open-source Implementations of trainable segmentation:

1) Trainable Segmentation Plug-in
2) IMMI Segmentation benchmarking.
   Several segmentation benchmarks are available for comparing the performance of segmentation methods with the state-of-the-art segmentation methods on standardized sets 1. Prague On-line Texture Segmentation Benchmark
3) The Berkeley Segmentation Dataset and Benchmark

V. Performance Evaluation

For the performance evaluation, we have chosen some typical algorithms i.e specifically ones which the authors used the Wallflower dataset to evaluate them. This dataset is the most used and consists in a set of image sequences where each sequence presents a different type of difficulty that a practical task may meet. The performance is evaluated against hand-segmented ground truth. Three terms are used in evaluation: False Positive (FP) is the number of background pixels that are wrongly marked as foreground; False Negative (FN) is the number of foreground pixels that are wrongly marked as background; Total Error (TE) is the sum of FP and FN. A brief description of the Wallflower image sequences can be made as follows:

1) Moved Object (MO) - A person enters into a room, makes a phone call, and leaves. The phone and the chair are left in a different position.
2) Time of Day (TOD) - The light in a room gradually changes from dark to bright. Then, a person enters the room and sits down.
3) Light Switch (LS) - A room scene begins with the lights on. Then a person enters the room and turns off the lights for a long period. Later, a person walks in the room, switches on the light, and moves the chair, while the door is closed.
4) Waving Trees (WT) - A tree is swaying and a person walks in front of the tree.
5) Camouflage (C) - A person walks in front of a monitor, which has rolling interference bars on the screen. The bars include similar color to the person's clothing.
6) Boostrapping (B) - The image sequence shows a busy cafeteria and each frame contains people.
7) Foreground Aperture (FA) - A person with uniformly colored shirt wakes up and begins to move slowly.

In Fig. 1, we have represented the overall performance for the five first algorithms and in Fig. 2 for the seven algorithms but without the image sequences Moved Object, Time of Day and Light Switch. Fig. 1 and Fig. 2 are not intended to be a definitive ranking of these algorithms. Such a ranking is necessarily task-, sequence-, and application dependent.

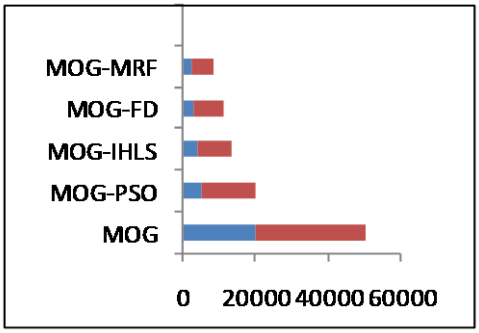

Fig. 1. Overall performance.

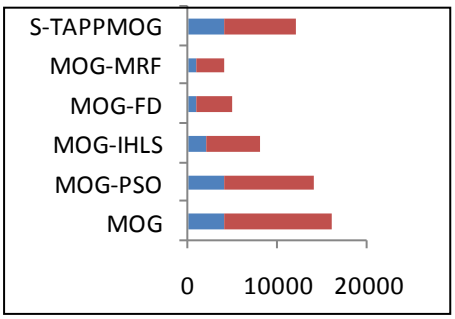

Fig. 2. Overall performance without MO, TD and LS.

VI. Conclusion

The various video segmentation algorithms are reviewed and the drawback of implementing the Gaussian mixture model on FPGA can be overcome by adapting the trainable segmentation to improve the processing capability of the overall system. Due to the real time processing reduces the analysis of long term effects due to changes in algorithms and parametric changes. The main bottleneck of image processing algorithms is the high memory requirements.

References

[1] M. Genovese and E. Napoli, "ASIC and FPGA Implementation of the Gaussian Mixture Model algorithm for Real Time Segmentation of High definition Video," *IEEE Trans. Very large scale integration (VLSI) systems*, 2013.

[2] C. S. Kamath and C. Robust, "Background subtraction with foreground validation for Urban Traffic Video," *J. Appl. Signal Proc. Special Issue on Advances in Intelligent Vision Systems: Methods and Applications (EURASIP 2005)*, New York, USA, vol. 14, pp. 2330-2340, 2005.

[3] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," *Int Conf on Computer Vision, (ICCV 1999)*, Corfu, Greece, pp. 255-261, September 1999.

[4] J. Carranza, C. Theobalt, M. Magnor, and H. Seidel, "Free-Viewpoint Video of Human Actors," *ACM Trans on Graphics*, vol. 22, no. 3, pp. 569-577, 2003.
[5] T. Horprasert, I. Haritaoglu, C. Wren, D. Harwood, L. Davis, and A. Entland, "Real-time 3D motion capture," *Workshop on Perceptual User Interfaces (PUI 1998)*, San Francisco, California, pp. 87-90, November 1998.
[6] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and tracking using voxel data," *Int J Comp Vision (IJCV 2003)*, pp. 199 -223, July 2003.
[7] El Baf F., T. Bouwmans, and B. Vachon, "Comparison of background subtraction methods for a multimedia learning space," *Int Conf. on Signal Processing and Multimedia (SIGMAP 2007)*, Barcelona, Spain, July 2007.
[8] A. Pande, A. Verma, and A. Mittal, "Network aware optimal resource allocation for e-learning Videos," *The 6th Int Conf. on mobile Learning*, Melbourne Australia, October 2007.
[9] J. Warren, "Unencumbered full body interaction in video games," thesis, MFA design and technology parsons school of design, New York, USA, April 2003.
[10] D. Semani, T. Bouwmans, C. Frélicot, and P. Courtellemont, "Automatic fish recognition in interactive live videos," in *Proc. of the IVRCIA 2002*, Orlando, Florida, USA, vol. 14, pp. 94-99, July 2002.
[11] M. Piccardi, "Background subtraction techniques: A review," in *Proc of the Int Conf on Systems, Man and Cybernetics (SMC 2004)*, The Hague, The Netherlands, pp. 3199-3104, October 2004.
[12] S. Elhabian, K. El-Sayed, and S. Ahmed, "Moving object detection in spatial domain using background removal techniques - State-of-Art," *Recent Pat on Comput Sci*, 2008, vol. 1, no. 1, pp. 32-54.
[13] B. Lee and M. Hedley, "Background estimation for video surveillance," *Image and vision computing New Zealand*, pp. 315-320, 2002.
[14] J. Zheng, Y. Wang, N. Nihan, and E. Hallenbeck, "Extracting roadway background image: A mode based approach," *J. Transport Res Report*, pp. 82-88, 2006.
[15] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans on Patt Anal Mach Intel (PAMI 1997)*, vol. 19, no. 7, pp. 780 -785, 1997.
[16] M. Sigari, N. Mozayani, and H. Pourreza, "Fuzzy running average and fuzzy background subtraction: concepts and application," *Int J. Comput Sci Network Security*, vol. 8, no. 2, 2008, pp. 138-143.
[17] R. Chang, T. Ghandi, and M. Trivedi, "Vision modules for a multi sensory bridge monitoring approach," *ITSC*, pp. 971-976, October 2004.
[18] F. Porikli and O. Tuzel, "Bayesian background modeling for foreground detection," *ACM Int Workshop on Video Surveillance and SensorNetworks (VSSN 2005)*, pp. 55-58, November 2005.
[19] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for realtime tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis.Pattern Recognit.*, vol. 2 Fort Collins, CO, USA, pp. 1–7, Jun. 1999.

R. S. Sabeenian is currently working as a professor in ECE Department in SonaCollege of Technology, Salem, Tamil Nadu, and India. He received his Bachelors in Engineering from Madras University and his Masters in Engineering in Communication Systems from Madurai Kamaraj University. He received his Ph.D. Degree from Anna University, Chennai in the year 2009 in the area of digital Image processing. He is currently heading the research group named Sona SIPRO (SONA Signal and Image PROcessing Research Centre) centre located at the Advanced Research Centre in Sona College of Technology, Salem. He has published more than 65 research papers in various International, National Journals and Conferences. He has also published around seven books. He is a reviewer for the journals of IET, UK and ACTA Press Singapore. He received the "Best Faculty Award" among Tamil Nadu, Karnataka and Kerala states for the year 2009 given by the Nehru Group of Institutions, Coimbatore and the "Best Innovative Project Award" from the Indian National Academy of Engineering, New Delhi for the year 2009 and "ISTE Rajarambapu Patil National Award" for Promising Engineering Teacher for Creative Work done in Technical Education for the year 2010 from ISTE.He has also received a Project Grant from the All India Council for Technical Education and Tamil Nadu State Council for Science and Technology, for carrying out research. He received two "Best Research Paper Awards" from Springer International Conference and IEEE International Conference in the year 2010.He was also awarded the IETE Biman Behari Sen Memorial National Award for outstanding contributions in the emerging areas of Electronics and Telecommunication with emphasis on R&D for the year 2011.The Award was given by Institution of Electronics and elecommunication Engineers (IETE), New Delhi.He is the Editor of 6 International Research Journals Research Journal of Information Technology, Asian Journal of Scientific Research, Journal of Artificial Intelligence, Singapore Journal of Scientific Research, International Journal of Manufacturing Systems and ICTACT Journal of Image Processing. He is also associated with the Image Processing Payload of the PESIT Pico Satellite Project which is to be launched by the end of Feb,2013.He is the External Expert Member for Board of Studies of Adhiyaman College of Engineering, Hosur and M.Kumarasamy College of Karur.He is the Honorary Treasurer of IETE Salem Sub Centre from 2010 onwards. He is the Co-ordinator for AICTE-INAE DVP Scheme. His areas of interest include texture analysis, texture classification and pattern recognition. He delivered more than 50 guest lectures and chaired more than 25 national and international conferences.