Proceedings of the ICEECE
December 22-24, Dhaka, Bangladesh

Automatic Video Object Segmentation from VOP

Ferdous Ahmed Sohel*[1], Chowdhury Mofizur Rahman** and Gour C. Karmakar***

*Dept. of CSE, International Islamic University Chittagong – Dhaka Campus.
** Dept. of CSE, United International University, Dhaka.
*** Gippsland School of Computing and Information Technology, Monash University, Australia.

ABSTRACT

The video coding standard MPEG-4 is enabling content-based functionalities of a prior decomposition of sequences into video object planes (VOP) so that each VOP represents a semantic object. Therefore extraction of semantic objects is an important part. There are various coding tools: shape coding, motion estimation and compensation, texture coding, multifunctional coding, error resilience, sprite coding and scalability. These are performed by using diverse techniques, like: pixel based segmentation, region based segmentation, boundary based segmentation, morphological segmentation, Bayesian segmentation, model based approaches etc. However, most of the techniques are either manual or semi-automatic. Semiautomatic methods require user assistance and suffer from addressing the following issues: the background is variable due to the camera motion, the light condition slightly changed, new object can appear at any time, objects may remain for a long time in the scene, many objects may be in a scene and possible occlusion. So it is very important to develop automatic techniques that are robust and fast and will combine low-level automatic feature segmentation with interactive method for defining and tracking high semantic video objects and also the above mentioned constraints. In this paper we have proposed how this might be done. It can be done in accordance with the following modules: sprite generation from background and statistical feature analysis approach for object definition, region based motion estimation for tracking.

1. INTRODUCTION

Successful video segmentation is necessary for most multimedia applications. New video coding standards, such as: MPEG-4 [1][2][4], MPEG-7 [3], do not only concentrate on efficient compression method but also concentrate on providing better ways to represent, integrate and exchange visual information. The MPEG-4 enables content-based functionalities by introducing the concept of video object planes. To take advantages of the object based functionalities defined in MPEG-4; a prior decomposition of video sequences into semantically meaningful object is required.

Recent development of semantic video object segmentation leads to three different types of algorithms:

1. Chromakey Scene Generation
2. Background subtraction and object motion tracking
3. Manual or semiautomatic

For semiautomatic algorithms, the user is required initially to identify the semantic objects interested. When the corresponding regions are passed to the computer these regions are tracked temporarily from the previous frame [5]. Since the temporary will introduce boundary errors, the region boundary needs to be modified and updated according to some low level homogeneity in the current frame.

Since human visual system (HVS) is very sensitive to the edge and contour information, the exact extraction of object boundaries is crucial for the success for object-based functionalities of video applications. Moreover, the semiautomatic or manual methods cannot accurately address to the following constraints: the background is variable due to the camera motion; the light condition slightly changed; new object can appear at any time; objects may remain for a long time in the scene, many objects may be in a scene; and possible occlusion. Some algorithms [6][7] have been devised. But these are not suitable enough to real time automatic applications. In this paper we have proposed an automatic video object segmentation procedure.

2. LITERATURE REVIEW

Video segmentation is the process of dividing a sequence of frames into smaller meaningful units that represent information at the scene level. Object segmentation has added some new dimensions in this field. An MPEG-4 video bit stream provides a hierarchical description of a visual scene as in figure 1. The hierarchical levels that describe the scene most directly are Visual Object Sequence (VS), Video Object (VO), Video Object Layer (VOL), Group of Video Object Planes (GOP) and Video Object Plane (VOP). VOP is a time sample of VO.

[1]Corresponding author E-mail: ferasohel27@yahoo.com

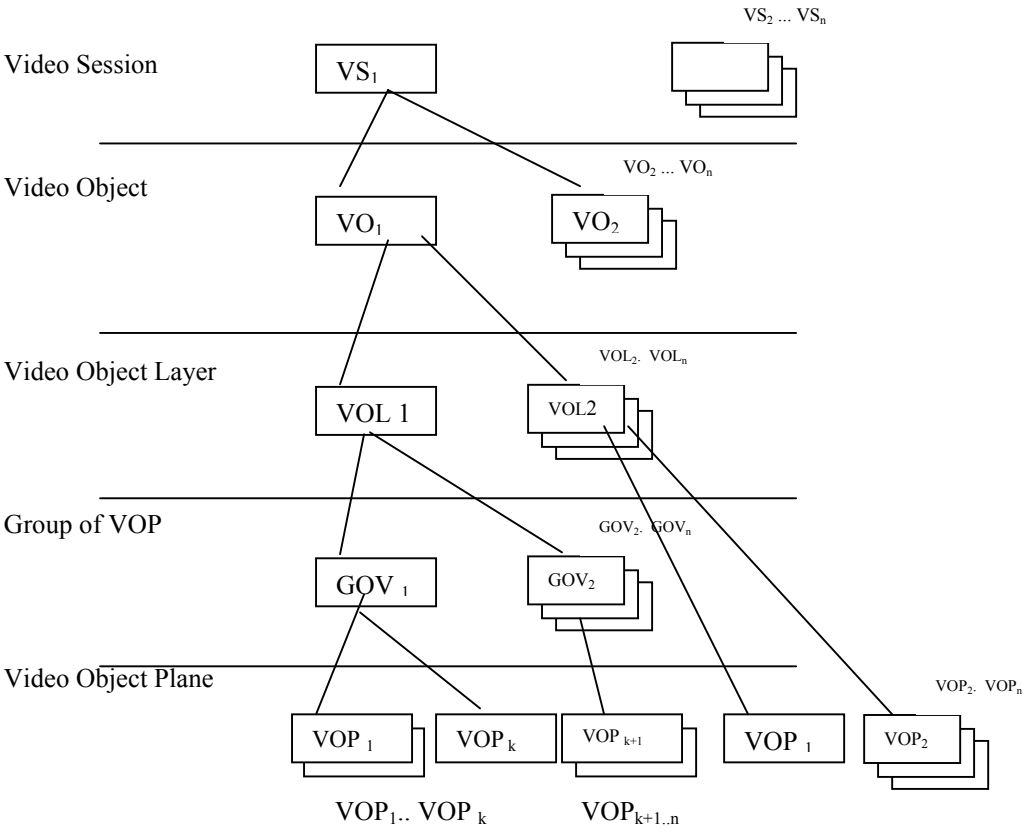

Figure 1: MPEG-4 video bit stream logical structure (C) 2000 ISO/IEC.

For MPEG-4 based coding it's essential to have the video object in advance to the encoding. However, most of the existing video clips are frame based. The video segmentation that aims at the exact separation of moving objects from the background becomes the foundation of the MPEG-4 object based video coding.

A lot of works have been done in the field of object segmentation and have ramified the field into various branches. Parametric motion models are used for temporal tracking [10]. Some other algorithms like watershed algorithm [9], and active contour "snake" algorithm etc are used for spatial region boundary updates [10, 11]. The boundary-based algorithms mainly depend on the detection of edge and ridge and valley. Since HVS is very sensitive to edge and contour, exact extraction is very much crucial in these aspects. Segmentation using estimation of relative depth [12] is very much influential now a days. Mosaicing is used to represent images with layers [13]. Some other algorithms work with Face Definition/ Animation Parameters (FDP/FAP) [14]. Besides, Sprite Generation [8] and transmission are very much influential techniques, which can use the adaptive techniques. AMOS [7] is also a very good such algorithm; but it is semiautomatic – it needs manual initialization. Various semiautomatic methods have been proposed for the segmentation of video sequences.

Pixel based: Segmenting objects based on multiple features such as color, intensity, texture and motion of the selected pixels, pixels are classified into one of the object types [15][16]. Also, an energy minimizing elastic contour model, which finds the best position according to the energy function, is used to track the moving contour of the object [17].

Region based: To find an object boundary, the user roughly marks the positions, which contains the real object boundaries. An automatic split and merging algorithm incorporates the user information in defining the object boundaries [5][10][18]. The extracted object can be used for object tracking for the subsequent frames.

But for semiautomatic algorithms, the user is required to initially to identify the semantic objects interested. This process may not address some crucial constraints mentioned in the introduction. So automatic segmentation is the state-of-art goal of current researchers.

3. METHODS

In this paper we have proposed automatic video object segmentation that will address to the above mentioned problems and serve real time automatic applications. In this approach there are two steps: initial sprite generation and tracking the moving objects. The block diagram for the approach is as in figure 2.

In the first step, a semantic object sprite will be generated adaptively. This generated object is then extracted through region partition and gradual region merging and bi-directional boundary refinement process. To obtain the precise semantic object boundary, the watershed detection method [9][10][11] can be used for region partition.

In the second step, successive frames are partitioned to region and every region is classified to foreground or background using backward projection techniques. In the proposed model the segmented objects will be semantically meaningful, will be spatially accurate, will be able to handle delete/introduction of objects.

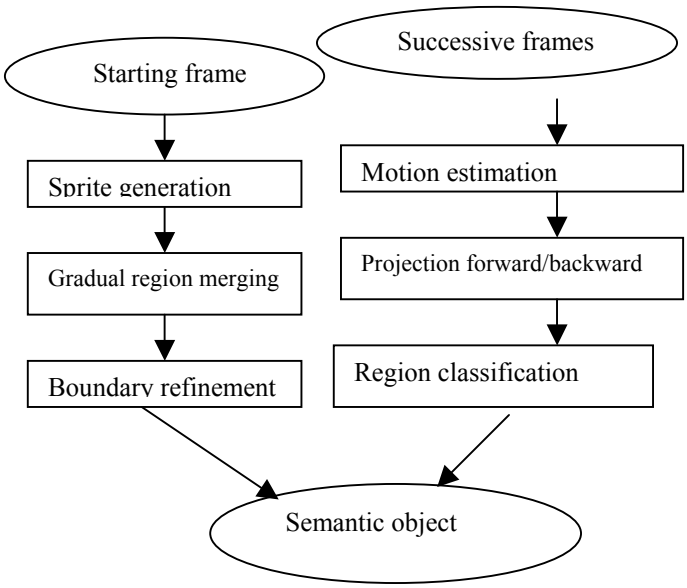

Figure 2: Block diagram of proposed methodologies.

Sprite generation: A sprite is an image composed of pixels belonging to a video object visible throughout a video sequence. For instance, in a panning sequence, portions of the background may not be visible in certain frames due to the occlusion of the foreground object and camera motion. The sprite generated will contain all the visible pixels of the background object through out the sequence. The video sequence consists of a few frames. These frames continually provide changed background information according to the camera panning. From the motion information, the sprite object background is generated and the reference frame can be obtained as a part of the sprite object. The reference frame extracted from a sprite object can be looked as a background, which can be used to easily identify the uncovered background area, which does not belong to a moving object. It is much

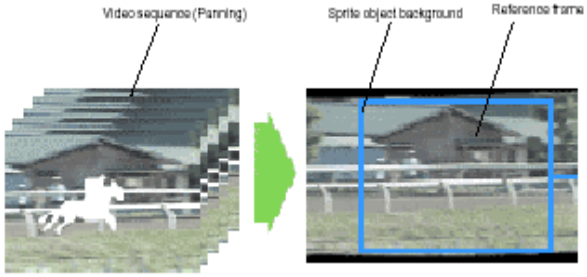

Figure 3: Sprite generation techniques.

easier to detect the overlap of two successive instances than using the simple subtraction of two successive frames. The methods of sprite object compression has been defined in MPEG-4 Verification Model (VM). However, the generation of sprite is an open issue. Global motion estimation (GME) is the main method of sprite generation. Hierarchical GME algorithms are very influential. This field can be further investigated. Adaptive background update control can be used to adopt the change events. Probability distribution or other statistical measures can be taken into account to acknowledge any change. The main change of the background of the video sequence can be camera break, background dissolve and camera motion such as zooming and panning. Adaptive concept is an important issue during the design of the framework. The adaptive reference frame update addresses many problems in other video segmentation approaches. For instance, the framework is suitable in the situations when the background change due to camera motion. Also it works in a light variable environment.

Initial region partition: The generated sprite is then simplified by using morphological filters in the YcbCr color space. And preserve the contour of the object. The spatial gradient of the simplified image is obtained using Sobel operator in the RGB color space. The watershed algorithm could be used to the spatial gradient image. It partitions the region into homogeneous groups.

We have used morphological opening/closing by reconstruction filters for image simplification with the boundary preservation. These filters pair removes the regions that are smaller than a given structuring element size but preserve the contours of the remaining objects in the image. These reconstruction filters applied to Y, Cb, Cr components separately.

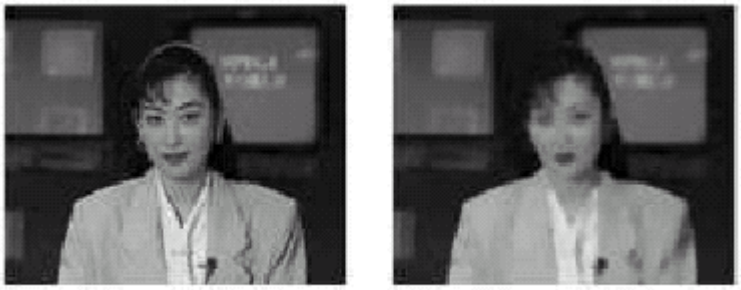

Figure 4: Image Simplification: a) Original Image b) Simplified image

The spatial gradient of the simplified image is obtained using a Sobel operator. The spatial gradient image is processed by the watershed algorithm which partitions image into homogeneous intensity regions. Figure 5 shows gradient images and region-based images in YCbCr and RGB color spaces.

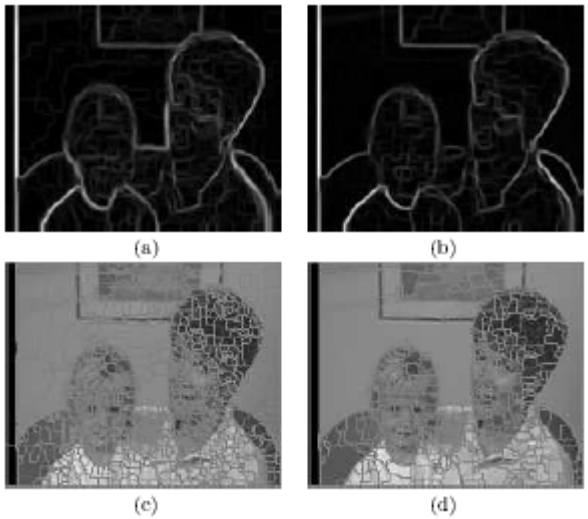

Figure 5: a) Gradient image in YCbCr color space b) Gradient image in RGB color space c) Watershed regions in YCbCr color space d) Watershed regions in RGB color space

The image is often interpreted as geographical surface in mathematical morphology and its grey level is treated as altitude. As shown in figure 6, the rain falling watershed algorithm is comprised of two steps. First, some of the weakest edges can be removed by "drowning" the image. The drowning step will create a number of "lake" grouping all the pixels that lie below a certain threshold. This useful to reduce the influence of noise and reduce over-segmentation. Second, for each pixel, we determine in which direction a raindrop would flow if it would fall on the topographic surface. This steepest decent neighbor and the pixel under consideration are then merged, finally enabling the localization of the remaining edges and segments, i. e., the areas surrounded by the topographic surface rims.

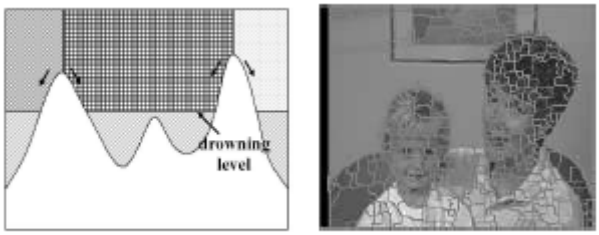

Figure 6: Water shed algorithm: a) Rainfall with drowning level, b) Watershed detection results.

We assume that watershed regions contain all the real object boundaries to be extracted. As shown in figure 7, the sprite, which is classified "foreground", "background" and "uncertainty" is superimposed on top of the spatial segmented regions obtained by watershed detection.

After labeling, an uncertainty region merging procedure for classifying to foreground and background is still required to obtain a real object boundary. The region similarity is then done by checking color similarity, edge strength and frame diffusion similarity. The measure is according to [5].

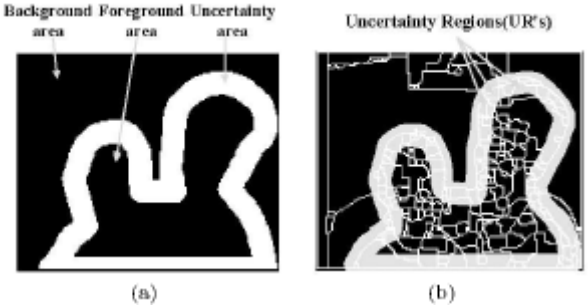

Figure 7: Region labeling: a) Sprite, b) Superimposed image and uncertainty regions.

Gradual Region Merging: To merge the over-segmented regions into semantic regions, a region-merging algorithm may be incorporated. In this portion fusing of color, edge information and frame differences are done. Regions may be labeled as foreground, background and uncertain afterwards. After labeling similarity for both spatial and temporal information may be done. Vector differences with likelihood functions [23] may be used to normalize the similarities.

Boundary refinement: Bidirectional boundary refinement algorithm may be used to remove errors introduced in the gradual region merging process to extract exact precise semantic object boundaries. Physics based boundary estimation may also be incorporated.

Inter-frame segmentation: It's the second part. To ensure the tracking of moving objects several steps may be done: motion estimation, projection and region classification.

Motion estimation: The motion vector may be obtained by the change in pixel intensity levels.

Projection forward/backward: To track the object in the subsequent frames forward tracking [19][20] method may be used. Forward tracking technique projects the previous segmentation result obtained at previous frame onto the current frame according to the motion information. In some cases forward tracking techniques may result more distortions [21]. So backward tracking [21][22] techniques may also be used. Backward tracking consists of three steps: region-based motion estimation, backward projection and region classification.

Region classification: Region classification determines whether each region in the current frame belongs to the semantic object or not. If maximum of the wrapped region lies inside the previous semantic object, the region is classified to be inside the currents semantic object.

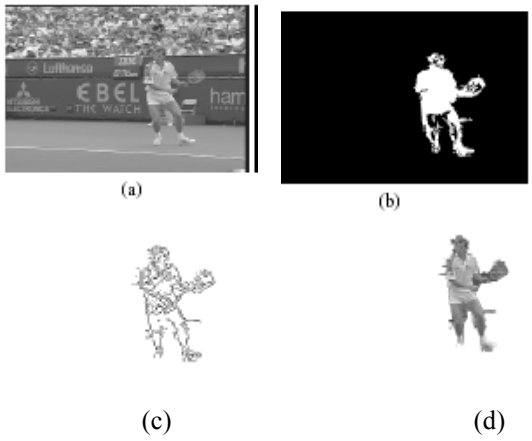

Figure 8: Object Segmentation: a) &b) the original objects in the video, c) &d) Extracted Objects respectively.

RESULTS AND CONCLUSION

In this paper, we have presented an automatic algorithm for object segmentation. It provides module-based approaches, which are used in real time applications. A very important issue presented in the paper is that, we develop the idea of getting the reference frames from sprite rather than user assistance. This approach can address all the previously identified problems in the introduction and it can be used in many practical situations. Automatic traffic controlling system, MPEG-4 based cartoons etc and many real time multimedia applications will be very much effective if robust, fast and automatic VOP extraction and object segmentation is devised. I hope that our proposed algorithm will be a small step towards it.

REFERENCES

[1] T. Sikora, "The MPEG-4 video standard verification model," *IEEE Trans. Circuits System Video Technology,* vol. 7, pp. 19-31, Feb. 1997.

[2] T. Ebrahimi, "MPEG-4 video verification model: A video encoding/decoding algorithm based on content representation," *Signal Processing: Image Communication,* vol. 9, pp. 367-384, 1997.

[3] MPEG-4, MPEG-7: Applications document, Tech. Report ISO/IEC JTC1/ WG11/W2860, MPEG, Vancouver, Canada, July, 1999.

[4] Official MPEG website, http://www.cselt.it/mpeg.

[5] Y. R. Kim, J. H. Kim, Y. Kim, S. J. Ko, "Semiautomatic segmentation using spatio-temporal gradual region merging for MPEG-4," *IEICE Trans. Fundamentals,* vol. E86 A, no. 10, October 2003.

[6] Corriea, Paulo; Pereira, Fernando; "Proposal for an Integrated Video analysis framework," ICIP 1998, Chicago, October 1998.

[7] Di Zhong and Shih-Fu Chang, "AMOS: An Active System for MPEG-4 Video Object segmentation", *IEEE International Conference on Image Processing,* October 4-7, 1998, Chicago, USA.

[8] Yan Lu, Wen Gao, Feng Wu, "Fast and Robust Sprite Generation for MPEG-4 Video Coding," *IEEE Pacific Rim Conference on Multimedia 2001*: 118-125.

[9] P. D. Smet, D. D. Vleeschauwer, "Performance and scalability of a highly ooptimized rainfalling watershed algorithm," in *Proceedings of CISST'98,* pp. 266-273, USA, July 1998.

[10] Chuang Gu and Ming Chieh Lee, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 8, pp. 572-584, September 1998, and reference therein.

[11] Munchurl . Kim, J. G. Choi, M. h. Lee, C. Ahn, "User assisted segmentation for moving objects of interest," *Doc. ISO/IEC JTC1/ WG11/M2803,* July 1997.

[12] M. Pardas, "Relative depth estimation and segmentation in monocular schemes," *Picture Coding Symposium,* PCS 97, Berlin, Germany, September 1997, pp. 367-372.

[13] J. Y. A. Wang, E. H. Adelson, "Representing Moving Images with Layers," *IEEE Trans. On Image Processing,* vol. 3, pp. 625-638, September 1994.

[14] M. J. T. Reinders, P. J. L. van Beek, B. Sankur, JCA van der Lubbe, "Facial Reature localization and Adaptation of a generic face model for model –based coding" *Signal proc.: Image Comm.,* 7(1995), pp. 57-74.

[15] E. Chalom, M. Bove, "Segmentation of an image sequence using multidimensional image attributes," in *Proceedings of International conference on Image Processing,* pp. 525-528, Lausanne, Switzerland, 1996.

[16] C. Toklu, A. M. Teklp, and A. T. Arden, "Semantic Video object segmentation in presence of occlusion," *IEEE Trans. On CSVT,* vol. 10, no. 4, pp. 624-629, June 2000.

[17] N. Ueda, K. Mase, "Tracking moving contours using energy minimizing elastic contour models," *Computer Vision-ECCV'92,* vol. 588,pp. 453-457, 1992.

[18] R. Castagno, T. Ebrahimi, M. Kunt, " Video segmentation based on multiple features for interactive multimedia applications," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 8, pp. 562-571, September 1998.

[19] F. Moscheni, F. Dufaux, M. Kunt, "Object tracking based on temporal and spatial information," in *Proc. ICASSP'86,* vol. 4, pp. 1914-1917, Atlanta, May 1996.

[20] F. Marques, C. Monila, "Object Tracking for content based functionalities," *VCIP'97,* vol. 3024, no. 1, pp. 190-199, San Jose, February 1997.

[21] Chuang Gu and Ming Chieh Lee, "Semantic video object tracking using region based classification," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 8, pp. 643-647, September 1998.

[22] D. Gatica Perez, M. Sun, C. Gu, "Semantic Video Object extraction based on backward tracking of multi-valued watershed ", in *ICPC' 99,* Kobe, Japan, October 1999.

[23] B. Sklar, Digital Communication, Prentice Hall, pp. 132-138, 1998.