Automatic Video Segmentation and Tracking for Content-based Applications

Hongliang Li, *Member, IEEE,* and King N. Ngan, *Fellow, IEEE*

The Chinese University of Hong Kong, Hong Kong SAR

{hlli, knngan}@ee.cuhk.edu.hk

Abstract—Advanced multimedia applications will have to provide content-related functionalities such as search and retrieval of meaningful objects, detection and analysis of events, and understanding of scenes, which allow the user to access and manipulate the multimedia content with greater flexibility. This greatly depends on automatic techniques for extracting such objects from multimedia data. In this article we intend to provide a tutorial on the state-of-the-art in video segmentation and tracking technology with particular attention paid to the recent developments in attention-based object extraction. Performance results are included to highlight this emerging technology.

I. Introduction

In the past several years, there has been rapid growing interest in content-based applications of video data, such as video retrieval and browsing, video summarization, video event analysis, video editing. From the content-related services, a semantic object (i.e., meaningful entity including a collection of attributes) can be detected and exploited to provide the user with the flexibility of content-based access and manipulation, such as fast indexing from video databases, advanced editing and composition, and efficient coding of regions of interest [1].

Video segmentation has been a key technique for semantic object extraction and plays an important role in digital video processing, pattern recognition, and computer vision. The task of segmenting/tracking a video object emerges in many applications, such as bank transactions monitoring, surveillance and video conferencing. A limited set of applications of video segmentation can be presented as follows:

- Video surveillance, where the segmentation result is used to allow the identification of an intruder or of an anomalous situation, and helps to anticipate and reveal patterns of their actions and interactions with one another in their environment to determine when "alerts" should be posted to security unit.

- Content-based video summarization, such as sports event summary, video skimming, video pattern mining, which requires the segmented semantic objects to perform the content classification, representation, or understanding.
- Content-based coding application in which each frame of a video sequence is segmented into semantically meaningful objects with arbitrary shape. This makes it possible to manipulate the object independently, so that suitable coding algorithm can be applied for each object resulting in subjective quality improvement.
- Computer vision, such as video matting, video tooning, and rendering, where the segmented 2-D objects from the input image or video sequences can be used for 3-D scene reconstruction.
- Videoconferencing and videophony applications, in which segmentation can achieve a better quality by coding the most relevant objects at higher quality.
- Digital entertainment, where some specific objects can be replaced by segmentation, such as the video games.

There are other possible applications, such as industrial inspection, environmental monitoring, or the association of metadata with the segmented objects, etc.

Unlike the non-semantic segmentation that aims to extract some uniform and homogeneous regions with respect to some texture or color properties, semantic video segmentation can be defined as a process which typically partitions the video images into meaningful objects according to some specified semantics. After video object extraction, the segmentation in the subsequent frames can be achieved by the object tracking scheme. There are many video segmentation methods in the current literature, which exploit different information, i.e., spatial, temporal, or spatio-temporal. The spatial information can be derived from edge or region by means of the measure of intensity changes, whilst the temporal information can be generated by a change detection technique over multiple frames. These methods can be simply summarized as the implementation of four phases: *object mask generation, post processing, object tracking, and object mask update*. The object mask is used to indicate the arbitrary shape of the video object where those pixels inside the object will have the corresponding object label. Certainly, the background can be viewed as a special object which may contain multiple regions that do not belong to the objects of interest. Once the mask of the object of interest is separated from the background of the current frame, the object in the subsequent frames can be identified and updated from frame to frame by the video tracking technology. The block diagram of a class of segmentation methods is depicted in Fig. 1.

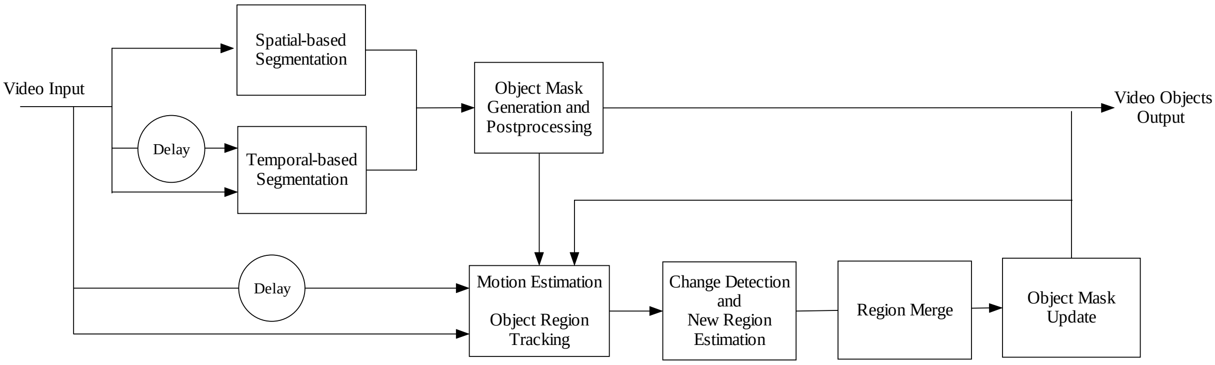

Video Input
Spatial-based Segmentation
Delay
Temporal-based Segmentation
Object Mask Generation and Postprocessing
Video Objects Output
Delay
Motion Estimation
Object Region Tracking
Change Detection and New Region Estimation
Region Merge
Object Mask Update

Figure 1. *Block diagram of the video segmentation and tracking algorithm.*

In the object mask generation process, the spatial and temporal information are usually employed in most video segmentation algorithms. Spatial segmentation is basically image segmentation, which partitions the frame into homogeneous regions with respect to their colors or intensities. This method can be typically divided into region-, boundary-, and classification-based approaches.

Region-based methods [2] rely on the spatial similarity in color, texture, and other pixel statistics to identify the 'homogeneity' of these localized features. Regions growing, merging and splitting techniques, which adopt a certain decision rule to test the homogeneity of a set of primitive regions, also belong to the same category. Boundary-based approaches use primarily a differentiation filter to detect the image gradient information and extract the edges. The discontinuous edges are then grouped to form the object contour. This approach can be employed to support region-based algorithms to improve the segmentation result. The main drawback of boundary-based approaches is their lack of robustness during the contour closure extraction because of the difficult computation of the region's closed boundaries. In the classification-based approach [3], a partition of the feature space is first created. The training and learning phases are then used for the classifier. This method enables a combination of cues, such as texture, color, and depth. In order to achieve high classification accuracy, non-linear decision functions are usually required when the sequences contain complicated content.

The spatial-based segmentation approach can provide more accurate object boundary than temporal-based method because of the high spatial correlation between the adjacent pixels within the object region. However, since relatively high computational complexity is involved, this method is usually applied to segment the first frame (the key frame), or the frames when the target objects are missed due to the fast object motion or corrupted in the presence of occlusions.

On the other hand, temporal segmentation, which is based on change detection followed by motion analysis, utilizes intensity changes produced by the motion of moving object to locate the position and boundary of objects in time and space. The change detection masks are the most common forms of motion information incorporated into the segmentation process, which can be represented with the absolute difference between two consecutive frames. Unlike the spatial segmentation approaches, higher efficiency can be achieved because of small number of operations for the segmented moving region instead of the whole image for every frame. However, lighting variation and noise might be incorrectly assigned to moving objects. It is usually very difficult to distinguish between changes due to true object motion and changes due to noise, shadow effects, etc.

In addition, many segmentation approaches based on graph cut and level set can be found in the literature. By treating image segmentation as a graph partitioning problem, graph cuts can be employed to find the globally optimal segmentation. For example, an active graph cuts approach to max-flow/min-cut problems is presented in [4], which can effectively use a good approximate solution (initial cut) that is often available in dynamic, hierarchical, and multi-label optimization problems in vision. Based on optimization by matting and graph-cut, many interactive segmentation approaches have also been developed in order to improve the segmentation performance. An object of interest in a complex environment can be extracted successfully at the cost of interactive effort on the part of the user. Generally, these methods can provide users with much better segmentation performance than automatic ways.

Recently, a learning-based segmentation method has been introduced in [5]. This method learns from unsegmented, cluttered images using a generative probabilistic model incorporating both shape model and bottom-up cues of color and edge. An iterative procedure allows refinement of each object's segmentation rather than making any hard decisions.

In order to satisfy the future content-based multimedia services, the segmentation of semantic objects corresponding to meaningful video objects in the real-world scenes is urgently required. Since objects of interest usually correspond to multiple regions, which may have very great spatial-temporal variations, it is usually difficult to segment these objects automatically without any primary criteria for segmentation. An intrinsic problem of the 'blind' segmentation algorithms, which have no contextual knowledge assumption regarding the object being segmented, is that objects of interest may not be homogeneous with respect to low-level features or usually change with the environmental factors, such as lighting conditions, etc.

In this article we intend to provide a contribution in this direction by describing the research activities performed on video segmentation and tracking for content-based multimedia services, with particular attention paid to the recent developments in attention-based object extraction. The remainder of this article is organized as follows. First, we describe the basic video segmentation approaches with lower semantic level. Then we introduce our research advances on the saliency model based video segmentation. Some simulation results will be shown for the presented approaches. Finally, we draw the conclusions.

II. General Overview of Video Segmentation and Tracking

It is generally known that video segmentation can be decomposed into two sub-problems, namely video object segmentation and video object tracking. In this section, a brief overview of the spatial and temporal segmentation techniques is described.

A. Spatial-based Video Segmentation

In the previous discussion, we pointed out that spatial segmentation can be classified into region-, boundary-, and classification-based approaches. In the following, we concentrate on the description of the region- based spatial segmentation method that is widely used in image and video segmentation. The general segmentation scheme [2] can be seen as the implementation of three major steps: *simplification, homogeneity estimation,* and *partition optimization*. For easier spatial segmentation, images are first simplified by removing irrelevant information, such as the complexity of textured areas or unwanted details. Typical simplification is achieved by employing a certain filter to smooth the original image while preserving the image boundaries.

Then the homogeneity estimation can be used to gather a set of pixels with the similar features (e.g., color, texture) into a single region. One can use two main types of homogeneity criteria, namely deterministic and probabilistic criteria to analyze the feature space. If the region data can be viewed as a realization of a certain model, the region is said to be homogeneous.

Finally, the possible over-partitions should be modified and updated to reach an optimal segmentation by the splitting and merging operations. One region can be merged into its neighboring regions based on the similarity measure. The process is repeated until no more regions can be merged according to a homogeneity criterion.

B. Moving Objects Segmentation

Moving objects usually consist of multiple regions with different colors, intensities or textures, which are commonly used for object tracking and surveillance. Generally, it is difficult to segment moving objects using spatial techniques unless an appropriate decision is achieved on the merging of all its regions to form a single object of higher semantic meaning. As an important video feature, motion can be viewed as a low-level semantic information, which can provide the otherwise missing semantic information in cases where uniform motion is expected, or if the moving objects don’t overlap. Thus, different segmentation strategies may be selected according to the segmentation scenario [6].

In order to extract the moving objects, efficient motion detection and matching methods are required especially in the case of scenes where change detection masks have been shown to be ineffective. We have presented a segmentation algorithm [1] that automatically extracts moving objects from a video sequence, which is depicted in Fig. 2.

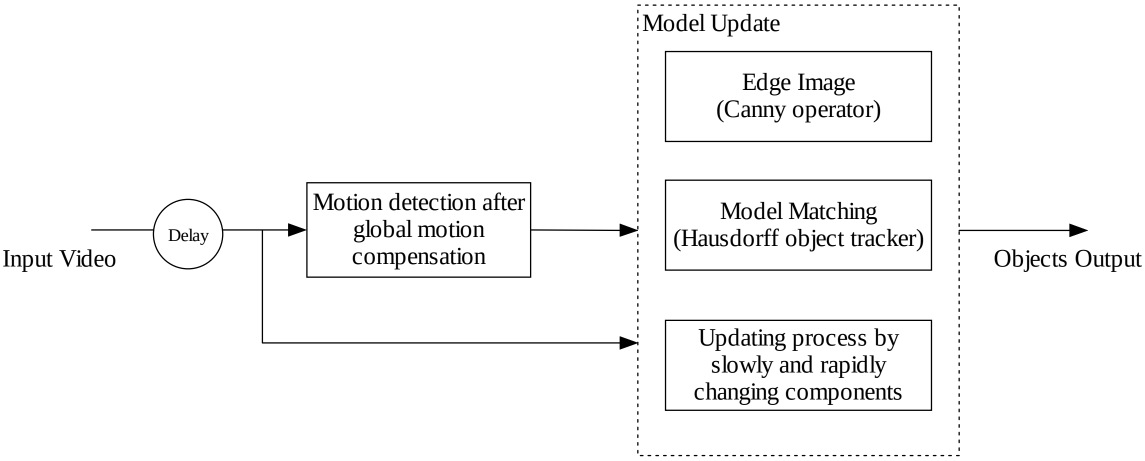

Model Update
Edge Image (Canny operator)
Input Video
Delay
Motion detection after global motion compensation
Model Matching (Hausdorff object tracker)
Objects Output
Updating process by slowly and rapidly changing components

Figure 2. *Block diagram of the moving object segmentation algorithm.*

The first step in this approach is to detect a dominant global motion that can be assigned to the background based on the six-parameter affine transformation. It is assumed that areas that do not follow this background motion show the presence of independently moving physical objects. To reduce the noise effect on the change detection, the connected components technique is employed to find those pixels belonging to the independently moving objects. If the size of a component exceeds a threshold, it will be classified to a moving object that can be used to obtain the initial object model by combining the edge pixels within a small distance.

In order to measure the temporal correspondence of objects and enhance the robustness to noise and changes in shape in the video sequence, an object tracker based on Hausdorff distance is established. This tracker holds a very useful property when dealing with objects that are partially occluded or rapidly changing their shape, and allows the user to keep track of objects even when they stop moving for an arbitrarily long time. For more information of this technique the reader is referred to [1].

Because of the possible changes in the object's shape, such as rotation or camera moving, the model must be updated every frame in order to handle the abrupt changes. The combination of two components, namely slowly changing parts and rapidly moving parts, is employed to achieve the robust updating mechanism. The first concentrates on the update of quasi-rigid parts that exhibit only small changes in successive frames, whilst the second is used to incorporate the non-rigid motion and newly appearing parts in the model update by adjusting the edge pixels.

A difficult task in this method is to obtain an initial model and to update the model of a non-rigid object with considerable changes in shape in the presence of a cluttered background. If the background edge points cannot be removed in the model matching and updating stages, the corrupted model with unreliable results is difficult to avoid. A hybrid strategy in the case of cluttered background can be found in [7] that aims to track object's region and to cope with the tracking management issues such as appearance and disappearance of objects, splitting and partial occlusions.

III. ATTENTION-BASED VIDEO SEGMENTATION

Visual attention is an effective simulation of human visual characteristic, which allows us to find relevant information quickly and efficiently. It has been successfully applied to many fields, such as pattern recognition, object detection and tracking, image/video coding, image database querying and retrieval, image adaptation, and video summary. The application in object extraction for color images is recently reported in [8], which is based on principal idea that human visual system only processes part of the incoming information in full detail, whereas the rest is left nearly unprocessed. Unlike the traditional methods, attention-based scheme aims to segment the meaningful physical entities that are more likely to attract viewers' attention than other objects in the video image. Most objects of interest in the video scene tend to be the attention objects that have distinctive features from their surroundings.

Based on the visual attention idea, we have constructed the saliency models to indicate the location of the objects in the video sequences. In order to develop an efficient and accurate approach in extracting interesting objects from video sequences in an unsupervised manner, the proposed algorithm uses the perceptual features that can be modeled by several low-level and high-level cues for the object of interest. First, we construct the saliency map based on the inherent features of the object of interest. Those features usually correspond to the object's specific properties, which can be obtained by the prior knowledge or the training procedure. Then, we employ non-linear filtering to eliminate the noise in the obtained saliency map, and use the classification approach to extract the object regions. The framework of our implementation approach is shown in Fig. 3.

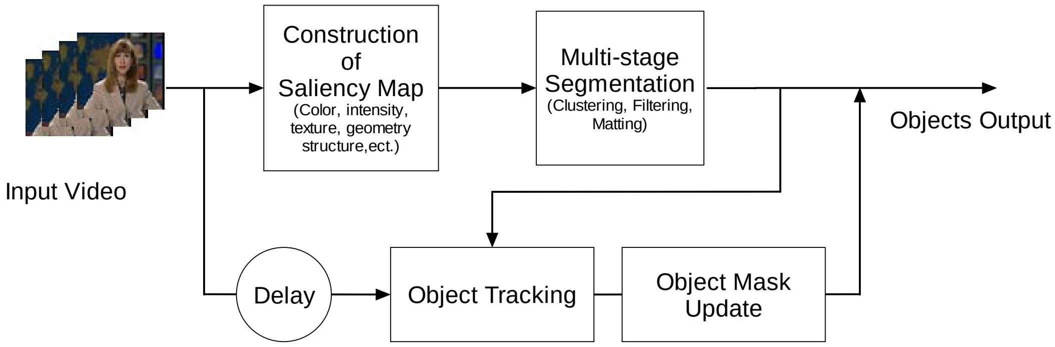
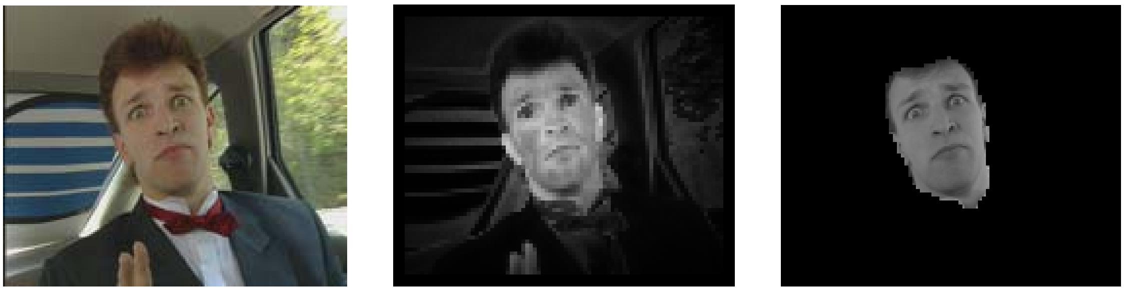

Construction of Saliency Map
(Color, intensity, texture, geometry structure,ect.)
Multi-stage Segmentation
(Clustering, Filtering, Matting)
Objects Output
Input Video
Delay
Object Tracking
Object Mask Update

Figure 3. *Framework of the implementation approach*.

Based on our framework, several saliency models are successfully constructed to extract the object of interest in different video sequences, such as the facial saliency model [9] and focused saliency model [10]. The first model can be regarded as the attention model based on the position, geometric structure, and color transformation, whilst the second model corresponds to the visual attention obtained from the analysis of cameras model. Generally, a typical camera is an optical system containing a lens and an image screen. When the object point is focused by the lens, a sharply point will be generated on the image plane whereas the far or the near object points, will appear as a circle of confusion. The focused saliency model of an image in our method is defined as the difference between the blurred and the original images that can capture the information of the focused object efficiently.

For the previous approach, we first generate the saliency map from the input head-and-shoulder type video image by our proposed facial attention model. Then, a geometric model and an eye-map built from the chrominance components are employed to localize the face region according to the saliency map. Here, the eye map is a binary image, which is determined by the chrominance distribution found around the eyes from a large number of training examples. The final step involves the adaptive boundary correction and the final face contour extraction. There are three conspicuity maps corresponding to the chrominance, luminance, and position information used to construct the facial saliency map. It is known that the face region usually exhibits similar skin-color feature regardless of different skin types [8]. The values of the chrominance component for the different facial skin colors are indeed narrowly distributed. Therefore, using the skin-color information, we can easily construct the facial saliency map to locate the potential face areas.

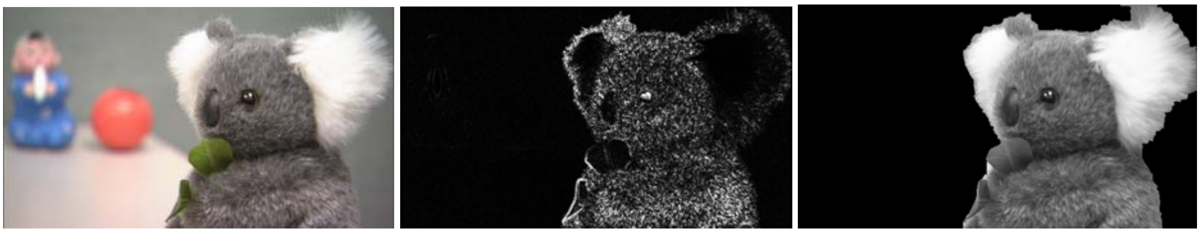
We have also found that in typical head-and-shoulder video sequences, most of the face locations appear at or near the center of the image in order to attract user attention. Few human faces are captured and shown at the boundary of the image, especially at the top or bottom of the image. In addition, although there is no narrow distribution in the face area for the luminance component, it is found that the darker the intensity value of a pixel, the less likely it will be a skin-tone color. Similar results can also be found for the very bright pixels. The reason is that a “head-and-shoulder” region, as the foreground, usually exhibits more uneven distribution of brightness, and provides a clearer visual result for user rather than the background. Experimental evaluation on test sequences shows that the proposed method is capable of segmenting the face area effectively, which is evident in Fig. 4.



Original Image Facial Saliency Map Segmented Result

Figure 4. *Segmentation based on facial saliency model*.

As mentioned above, the second saliency model is based on the camera model that extracts the object of interest by measuring the focused region. When any point is not in focus, it will appear as a blurred circle. The amount of defocus or blurring actually depends on the characteristics of the lens system and the distance to the surface of exact focus. Generally, only the object of interest is in sharp focus, whereas background objects are typically blurred or out-of-focus. Therefore, we first generate a re-blurred version of the input video image by a point-spread function in the proposed method. The focused saliency map of an image in our method can be obtained from the difference between the original and the blurred images. We can see that most of the energy in the saliency map corresponds to the focused object, whilst a large amount of the energy of the defocused region is removed efficiently. This feature is very useful in segmentation for it provides us with sufficient information about the segmented objects. Then, a bilateral filter and morphological operator are employed to smooth and merge the focused regions. The third stage involves Poisson matting approach to extract the focused region accurately from the obtained trimap. An example of the segmented result can be found in Fig. 5.



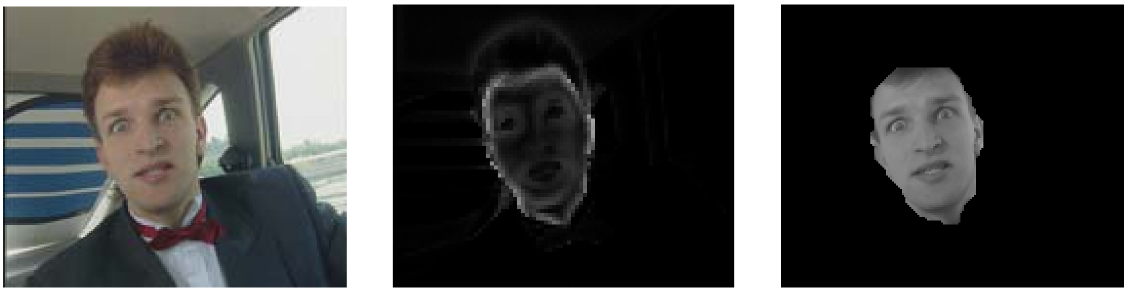
Original Image Focused Saliency Map Segmented Result

Figure 5. *Segmentation based on focused saliency model*.

The segmentation in the successive frames is achieved by the tracking procedure. The major steps in this stage are the boundary matching and connection. In our proposed method, a boundary saliency model is first constructed to determine the object boundary. Then a connective technique between two key points is employed to extract this initial region. As similar to our previous work in [1], both of the tracking schemes are based on the object boundary information. Apart from the edge information used in [1], we also incorporate the color and position information into the tracking procedure.

The first step for tracking the object region is the projection of the information in the previous frame onto the current frame. By applying the motion estimation technique, we can easily obtain the current position of each candidate object area. We use three “conspicuity maps” corresponding to the color, edge, and position components to construct the boundary saliency model. In this model, it can be observed that if the edge points that are close to the projected boundaries have the same color feature as the segmented object region in the previous frame, large boundary saliency values will be observed for these points. On the contrary, the background points tend to have smaller saliency values due to the inconspicuous color or unlikely position features.

In order to find the boundary points in the current frame and segment the corresponding object region, we use two stages to realize the extraction process. The first is the boundary matching, which aims to find the best points with maximal boundary saliency values for the projected object boundary. The second is the post-processing stage, which is used to connect two broken boundary points. This procedure is only applied to the disconnected contour after the boundary matching. It is known that if the deformation of the object contour appears during the object movement, such as the object approaching the camera, some new contour points will be generated, that causes the discontinuity of the matched boundary. Here, a simple linear connection method is employed to link the two broken boundary points to form a closed loop for the object contour. Finally, we use a filling-in technique to extract the corresponding area of interest. An example of the segmented results using the *carphone* sequence is illustrated in Fig. 6. Generally, a better segmentation result can be obtained with less computational complexity, which can be applied to video conferencing and videophony. On the other hand, it should be noted that since the simple linear contour interpolation has been employed, the segmentation errors will result especially in the case of fast and large shape changes for the segmented object. To obtain accurate segmentation results, we can use other optimization algorithm to improve the performance, such as matting or graph cut.



Original Image Boundary Saliency Map Segmented Result

Figure 6. *Segmentation based on boundary saliency model.*

IV. Conclusions

In this article, a survey of the different object segmentation schemes has been presented, making comparisons and classifying them into different categories. We have described the proposed segmentation algorithm that automatically extracts moving objects from a video sequence with particular attention paid to our recent work on the saliency model based video segmentation scheme, which aims to segment the semantically meaningful objects that are more likely to attract viewer's attention than other objects in the video image.

We have shown that video segmentation is an important technique which plays an important role in many fields, such as video processing, pattern recognition, and computer vision. The rapid growth of content-based multimedia applications will continue to attract more and more attention to this research task in the successful delivery of contents to users. However, it has been observed that it is hard to provide a uniform video segmentation solution because what is interesting and what is not depends on the application scenarios. More work needs to be done in researching the segmentation schemes for meaningful semantic objects (conventional schemes may not be suitable for this case). Furthermore, much research effort also needs to devote to segmentation quality evaluation, which allows the appropriate selection of segmentation algorithms in order to achieve optimal performance.

In addition, attention model as an efficient tool in visual signal processing allows us to quickly select the useful information that is relevant to the segmented objects. By combining multiple image features into a saliency map, the attended locations can be detected from the obtained saliency values. On the other hand, the modeling scheme of the visual attention needs further studies, because it is still a challenging problem to know what information is important enough to capture attention for the segmented object. In addition, more semantic features should be considered when the attended object shares common properties with the background or other objects.

ACKNOWLEDGMENTS

The work was partially supported by a research grants from the Shun Hing Institute of Advanced Engineering, and the Research Grants Council of the Hong Kong SAR under Project CUHK4229/04E. The authors would like to thank all the anonymous reviewers for their constructive comments and useful suggestions.

REFERENCES

[1] T. Meier and K.N. Ngan, "Automatic segmentation of moving objects for video objects plane generation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, no. 5, September 1998, pp. 525-538.

[2] P. Salembier and F. Marques, “Region-based representations of image and video: segmentation tools for multimedia services,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 8, December 1999, pp. 1147-1169.

[3] Y. Liu, and Y. F. Zheng, “Video object segmentation and tracking using ψ-learning classification,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no.7, July 2005, pp. 885-899.

[4] O. Juan, and Y. Boykov, “Active graph cuts, ” In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 1, pp. 1023 - 1029, June 2006.

[5] J. Winn, N. Jojic, “LOCUS: learning object classes with unsupervised segmentation,” *IEEE International Conference on Computer Vision* (*ICCV’ 05*), vol.1, pp.756 – 763, Oct. 2005.

[6] P. L. Correia and F. Pereira, “Classification of video segmentation application scenarios,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 5, May 2004, pp. 735-741.

[7] A. Cavallaro, O. Steiger, and T. Ebrahimi, “Tracking video objects in cluttered background,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no. 4, April 2005, pp. 575-584.

[8] J. Han, King N. Ngan, M. Li, and H.-J. Zhang, “Unsupervised extraction of visual attention objects in color images,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 16, no. 1, January 2006, pp. 141-145.

[9] H. Li, and King N. Ngan, “Face segmentation in head-and-shoulder video sequence based on facial saliency map”, *IEEE International Symposium on Circuits and Systems*, Kos, Greece, May 2006.

[10] H. Li, and King N. Ngan, “Unsupervised segmentation of defocused video based on matting model”, *IEEE International Conference on Image Processing*, Atlanta, GA, USA, Oct. 2006.