Probabilistic Spatio-Temporal Video Object Segmentation Incorporating Shape Information

Rakib Ahmed, Gour C. Karmakar and Laurence S. Dooley

Gippsland School of Information Technology
Monash University, Australia.
{Rakib.Ahmed, Gour.Karmakar, Laurence.Dooley}@infotech.monash.edu.au

ABSTRACT

Automatic segmentation of semantic video object is one of the most imperative and demanding research areas in recent years. Generally probabilistic spatio-temporal video object segmentation techniques consider only pixel features, which tends to limit their performance in segmenting arbitrary shaped objects. To address this limitation requires a strategy to embed shape information seamlessly into the segmentation process and this paper presents a new generic shape-based probabilistic spatio-temporal algorithm for segmenting video objects. Experimental results using a number of standard video test sequences reveal a significant performance improvement in being able to segment arbitrary shaped video objects in comparison with other contemporary and widely used space-time based video segmentation methods.

1. INTRODUCTION

Semantic video object segmentation is one of the most important and challenging contemporary issues in the literature because while humans can differentiate video objects effortlessly, computer-based fully automatic video object segmentation techniques still remain an intractable research topic in the multimedia technology field. It traverses many application domains from security to medical imaging, with its major areas, being, though by no means limited to, surveillance and object tracking, content based video retrieval and analysis, video footage analysis for various investigation purposes, traffic systems, video coding and medical diagnosis.

Video object segmentation algorithms can be broadly classified into three major categories: i) segmentation with spatial priority, ii) segmentation with temporal priority and iii) joint spatial and temporal segmentation [1]. In contrast to the first two classes which give priority to either spatial or temporal grouping of pixels, the third class considers any video sequence as a spatio-temporal block of pixels. The advantage of undertaking the processing in the joint spatial and temporal domain is that it exploits the complementary nature of spatial and temporal information. This type of video segmentation technique is supported by psychologists who have long recognised the human visual system often finds salient structures jointly in space and time [2], hence its interest to the relevant research community.

One of the most popular spatio-temporal video segmentation techniques is *probabilistic space-time video object* segmentation, which has a strong theoretical basis with the task of segmentation being formulated in a statistical probabilistic framework. In [3] a probabilistic space-time (PST) video segmentation method using a piecewise Gaussian mixture model (GMM) was proposed, which mapped a video sequence into six dimensional feature vectors consisting of space, colour and time. The feature vectors are characterised by the GMM with parameter estimation achieved using the well established *expectation maximization* (EM) algorithm [4]. The key feature of this technique is that it analyses video frames as a single entity for model estimation purposes, so a block of frames (BOF) with certain overlaps is considered and model estimation is performed within each individual BOF, under the assumption that the motion is approximately linear. While the approach has been widely used, it has the fundamental drawback of being very dependent on the motion and pixel features in each frame. Also the computational complexity increases as a result of considering BOF overlaps.

Colour and spatial location are other important features for object representation, though they are insufficient to represent all types of objects, as there are a huge number of objects and a myriad of variations amongst them. For this reason, in many cases colour and spatial features alone, fail to approximate objects precisely. This motivates consideration to visual attributes that more intrinsically represent an object, of which the most important perceptual attribute of any object is shape, as this provides valuable clues for humans in both distinguishing and recognising

objects. To date, while embedding generic shape-based information about non-rigid objects into a video object segmentation framework has not been considered, there have been recent examples of integrating elliptical shape detection in video object tracking [5], though most natural objects in a video sequence tend to be non-elliptic in shape.

This paper proposes a new probabilistic spatio-temporal video object segmentation algorithm that incorporates generic shape information into the foundation of PST [3]. In order to achieve efficient and accurate segmentation, the proposed technique automatically extracts and incorporates shape information in the GMM model using a number of chord lengths [6] passing through the centre of each object. The subjective performance of the new method has been evaluated and compared with [3] using a number of standard video test sequences including *Salesman, Carphone and Akiyo.*

The remainder of this paper is as follows: In Section 2 the theoretical underpinning of the probabilistic space-time video object segmentation technique is briefly outlined, while the fundamental theory in representing generic shape information and then integrating it into PST is detailed in Section 3. An analysis of the experimental results is given in Section 4, with some concluding remarks given in Section 5.

2. PROBABILISTIC VIDEO-OBJECT SEGMENTATION

In this algorithm, each pixel is represented by a six dimensional feature vector of space, colour and time. The *Lab* color space is used to characterize the pixel colours as it is shown [7] to be approximately uniform in perception and the distances in this colour space meaningful.

If the distribution of a random variable $X \in R^d$ is a mixture of *k* Gaussians, the density function is defined as,

$$f\left(x_i|\theta\right)=\sum_{j=1}^{k}\alpha_j \frac{1}{\sqrt{(2\pi)^d\left|\Sigma_j^{-1}\right|}} e^{-\frac{1}{2}\left(x_i-\mu_{ij}\right)^T\Sigma_{ij}^{-1}\left(x_i-\mu_{ij}\right)} \quad (1)$$

where the parameter set $\theta=\left\{\alpha_j,\mu_j,\Sigma_j\right\}_{j=1}^{k}$ in which $\alpha_j>0,\sum_{j=1}^{k}\alpha_j=1$; $\mu_j\in R^d$ and Σ_j is a $d\times d$ positive definite matrix. The maximum likelihood (ML) estimation of θ for a set of feature vectors $x_1,...,x_n$ is given by

$$\theta_{ML}=\arg\max_{\theta} L\left(\theta|x_1,...,x_n\right)=\arg\max_{\theta}\sum_{i=1}^{n}\log f\left(x_i|\theta\right) \quad (2)$$

The EM algorithm [4] is applied for estimation of parameters θ_{ML} for GMM. The EM algorithm is initialized by the *K*-means algorithm and iteratively obtains θ_{ML} using the following equations:

$$p_{ij}=\frac{\alpha_j f\left(x_i|\mu_j,\Sigma_j\right)}{\sum_{c=1}^{k}\alpha_c f\left(x_i|\mu_c,\Sigma_c\right)} \quad (3)$$

$$\hat{\alpha}_j \longleftarrow \frac{1}{n}\sum_{i=1}^{n}p_{ij} \qquad \hat{\mu}_j \longleftarrow \frac{\sum_{i=1}^{n}p_{ij}x_i}{\sum_{i=1}^{n}p_{ij}} \qquad \hat{\Sigma}_j \longleftarrow \frac{\sum_{i=1}^{n}p_{ij}\left(x_i-\hat{\mu}_j\right)\left(x_i-\hat{\mu}_j\right)^T}{\sum_{i=1}^{n}p_{ij}}$$

The information-theoretic framework based on the principle of Minimum description length (MDL) [8] is employed for model selection.

3. INCORPORATING GENERIC SHAPE-BASED INFORMATION

As alluded above, shape is one of the most recognisable properties of any object, so in the new algorithm, a probabilistic function for determining the likelihood of a pixel belonging to a certain shape (cluster) is introduced. One method for video object tracking employed elliptical shape information in terms of the major and minor axes in a maximum *a posteriori* (MAP) framework [4], however it was unable to consider generic shape information. In the new technique, this fundamental concept is extended to facilitate the representation of generic shape information, using a strategy involving a series of chords passing through the centre of each object. This approximates the shape of an object as the chords are produced by considering shape boundary points and chord length distribution then approximates the shape [6]. The technique to represent a generic shape is explained further in the following section.

3.1. Representation of generic shape information

The shape of each object is determined using the probabilistic segmentation algorithm [3] considering space and colour features described in Section 2. From the extracted shape contour, the major axis is determined and *n* number of chords generated at equal angular distances. Rotation and translation of objects are normalized by the estimation of the gradient of the objects' major axes and updating the object's centre respectively.

If an object to be segmented in frame *t* is assumed as an object layer *j*, the prior function for a pixel x_i belonging to layer *j* is defined as [4],

$$O_{tj}\left(x_i\right)=\gamma+e^{-\frac{1}{2}\left(x_i-\mu_{tj}\right)^T\Sigma_{tj}^{-1}\left(x_i-\mu_{tj}\right)} \quad (4)$$

where γ represents the uncertainty of the layer shape and μ_{tj} is the translation parameter. The covariance matrix Σ_{tj} is defined as

$$\Sigma_{tj}=R^T\left(-\omega_{tj}\right)Diag\left[l_1^2,...,l_n^2\right]R\left(-\omega_{tj}\right) \quad (5)$$

where $l_1,...,l_n$ are *n* chord lengths as shown in Fig.1 and ω_{tj}

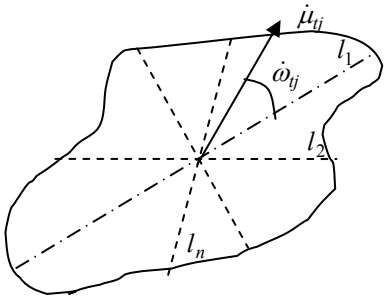


Fig.1: Illustration of shape using chords with rotation and translation parameters.

is the rotation angle. The longest chord through the centre of the object is termed the major axis. Using the contour points on the object boundary other chords are drawn through the centre at equal angular distances.

3.2. Pixel labelling

The joint probability for space, color and shape based probabilistic estimation of pixel x_i to be affiliated in layer j can be defined as,

$$S(x_i|w_t) = w_t f(x_i) + (1 - w_t) O_{tj}(x_i) \quad (6)$$

where w_t is the parameter that trades off between shape and spatio-color space, whose value is automatically determined by maximizing the probability function defined in (6). The labelling (hard decision) of each pixel is chosen as the maximum *a posteriori* probability given by:

$$L(x_i) = \arg\max_j S_{ij}(x_i) \quad (7)$$

and the confidence level (soft decision) of a particular pixel x_i belongs to cluster j is defined as:

$$P(L(x_i) = j) = S(x_i) / \sum_{j=1}^{k} S(x_i) \quad (8)$$

3.1. Locating the object of interest

Unlike user interactive video segmentation techniques, fully automatic approaches have to rely on certain parameters such as color, motion to extract semantically meaningful objects of interest. For real world practical applications, such as video surveillance, tracking, separating background, motion is the key feature for semantic object detection so therefore in the proposed algorithm, the correlation between the pixels belonging to a certain cluster in temporal direction is calculated. The correlation coefficient $R, (-1 \le R \le 1)$ is determined in order to find the motion of the segmented region, as well as the direction of motion [3]. Using the sign and magnitude of the correlation coefficient, the object (cluster) of interest can be determined throughout the video sequence considering its translation, rotation and shape information.

A key feature of the new technique is that the shape information is incorporated probabilistically which impacts on the probability of pixels to be labelled or assigned to a particular cluster. Explicit consideration of shape information also performs translation and rotation normalization for each object in a frame, which crucially then removes the requirement of having to take overlapping BOFs in order to find a correspondence between every pair of adjacent BOFs [3]. This advantage means the new algorithm reduces the amount redundant information that has to be processed during object segmentation.

The various steps involved in the new probabilistic shape-based video segmentation method are summarised in Algorithm 1.

Algorithm 1: *Probabilistic shape-based video segmentation*

Precondition: Video test sequence
Post condition: Segmented video object sequence.

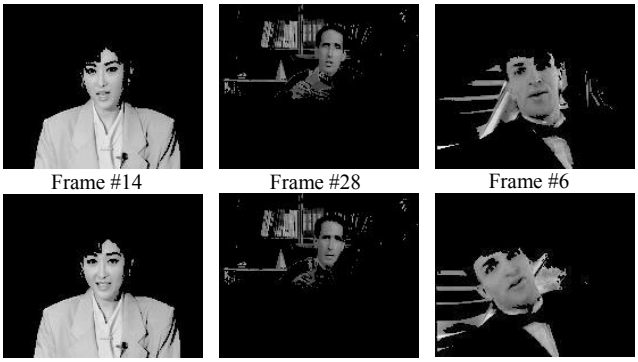
1. Extract feature vectors from a video frame and initialise GMM model parameters (1) using K-means algorithm.
2. Apply EM algorithm (3) to estimate GMM model parameters in **Step 1**.
3. Select model using the MDL principle.
4. Determine object shape from the clusters and define shape of a particular cluster with a specified number of chords passing through its centre.
5. Find major axis and centre of the cluster.
6. Calculate the joint probability of an object by (6).
7. Label each pixel using (7).
8. For event detection, determine the object and direction of motion using correlation coefficient.
9. STOP.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The new video segmentation approach has been implemented using MATLAB 7.0.1 running on Pentium-IV, 2.4 GHz CPU with 1 GB of memory. Experiments were conducted using three true colour QCIF standard video test sequences of frame-size 176×144 pixels. Fig.2. shows the representative examples and their respective frame numbers for the *Akiyo*, *Salesman* and *Carphone* video sequences, which have been used extensively to evaluate the video object segmentation performance in [1] and [9]. The *Akiyo* sequence does not contain any global motion but does have some object motion (newsreader talking). The *Salesman* sequence also does not possess any global motion, but this time the motion of the non-rigid object (salesman) is greater, especially in respect to the movement of his arms. Finally, unlike the other sequences, the *Carphone* sequence has both global and object motion, due to the fast moving

(a) Akiyo #32 (b) Salesman #28 (b) Carphone #88

Fig.2. Original video frames.

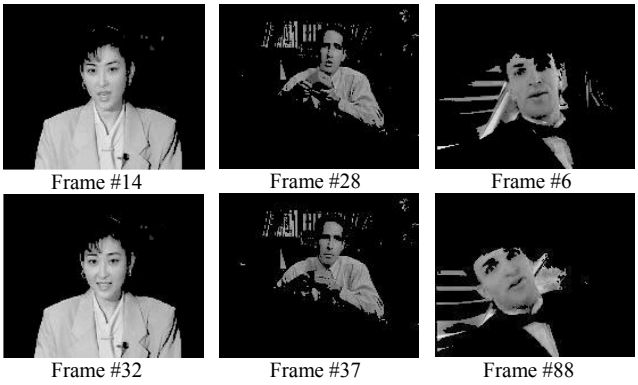


Frame #14 Frame #28 Frame #6

Frame #32 Frame #37 Frame #88

(a) Akiyo (b) Salesman (b) Carphone

Fig. 3. Segmentation using the PST technique.



Frame #14 Frame #28 Frame #6

Frame #32 Frame #37 Frame #88

(a) Akiyo (b) Salesman (b) Carphone

Fig. 4. Segmentation using the proposed technique.

background visible through the car window and object motion caused by the movement of the passenger.

The segmentation results for particular frames from the *Akiyo, Salesman* and *Carphone* sequences produced by the probabilistic space-time (PST) and the new segmentation technique are shown in Figs. 3 and 4 respectively. If the results for frame #32 of the *Akiyo* in Fig. 3 (a) and 4(a) are compared with the original frame in Fig. 2(a), it is visually apparent that a number of pixels in both the forehead and neck of *Akiyo* have been accurately separated by applying the new shape-based segmentation strategy, while the corresponding results for the *Salesman* in Fig. 4(b) show that a considerable portion of the body has been correctly segmented and the background region reduced when compared with the results produced by PST in Fig. 3(b), so vindicating the incorporation of generic shape information into the segmentation framework.

A similar trend is observed in the results for the *Carphone* sequence in Fig. 4(c) especially for the forehead and the chin of the person and correction of some scattered spots, despite this time the presence of global motion. It can be concluded from the results for three standard video sequences that shape integration has improved the segmentation of video objects over PST through out the frames of the video sequences.

6. CONCLUSIONS

Automatic segmentation of semantic video objects is a very challenging research topic, with most existing techniques being based upon pixel features, so they do not perform well for all video object types. This paper has introduced a new video object segmentation technique that seamlessly incorporates generic shape information about objects in a video frame sequence into a probabilistic spatio-temporal segmentation framework. Experimental results upon a number of different video test sequences have illustrated both the efficacy and benefit that integrating generic object-based shape information consistently provides in terms of improving the overall segmentation performance.

7. REFERENCES

[1] L. Liu, and G. Fan, "Combined Key-Frame Extraction and Object-Based Video Segmentation," IEEE Trans. on Circuits and Systems for Video Technology, vol. 15, no. 7, pp. 869-884, July 2005.

[2] R. Megret and D. DeMenthon, "A Survey of Spatio-Temporal Grouping Techniques," LAMP, CS-TR-4403, Univ. of Maryland, August 2002.

[3] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic Space-Time Video Modeling via Piecewise GMM," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 26, no. 3, pp. 384-396, March 2004.

[4] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood From Incomplete Data via the EM Algorithm," J. Royal Statistical Soc. B, vol. 39, no. 1, pp. 1-38, 1997.

[5] H. Tao, H. S. Sawhney, and R. Kumar, "Object Tracking with Bayesian Estimation of Dynamic Layer Representations", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 75-89, January 2002.

[6] S. P. Smith and A. K. Jain, "Chord Distributions of Shape Matching", Computer Graphics and Image Processing," vol. 20, pp-259-271, 1982.

[7] G. Wyszecki and W. Stiles, "Color Science: Concepts and Methods, Quantitative Data and Formulae," Wiley, 1982.

[8] P. D. Grünwald, I. J. Myung and M. A. Pitt, "Advances in Minimum Description Length Theory and Applications," The MIT Press, 2005.

[9] X. Song and G. Fan, "Key-Frame Extraction For Object-Based Video Segmentation," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, USA, March 2005.