Region-Based Shape Incorporation for Probabilistic Spatio-Temporal Video Object Segmentation

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

Embedding generic shape information into probabilistic spatio-temporal video object segmentation is of pivotal importance to achieving better segmentation, since it provides valuable perceptual clues for humans in both distinguishing and recognising objects. Recently a probabilistic spatio-temporal video object segmentation algorithm incorporating shape information has been proposed, though since it is restricted to only pixel features, the probability of a pixel belonging to a certain cluster is directly correlated with its spatial location, which theoretically limits the segmentation performance of the technique. To address this problem, this paper proposes a new probabilistic spatio-temporal video object segmentation algorithm that incorporates generic shape information based on its region. Experimental results reveal a significant performance improvement in arbitrary-shaped video object segmentation compared with other contemporary methods for a variety of standard video test sequences.

Index Terms— Image sequence analysis, object detection, shape, machine vision.

1. INTRODUCTION

Multimedia information is of paramount importance in order to perceive, recognize and understand any object with clarity and simplicity. With the incredible advances in hardware technologies and evermore powerful computers, it is more than ever becoming possible for multimedia information to be used in various aspects of our daily lives. Semantic video object segmentation is one of the most important and challenging issues for the breakthrough in multimedia technologies as 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. While the human eye can differentiate video objects effortlessly, fully automatic computer-based video object segmentation still remains a very challenging problem faced by the multimedia research community.

The importance of shape for video objects was initially recognised by the broadcast and movie industry in applying the well known *chroma-keying* technique. Algorithms such as object-based analysis-synthesis coding [1] use shape as a parameter together with texture and motion for describing moving video objects, while second-generation image coding segments an image into regions, describing each region using texture and shape [2]. The advantages of using shape include, achieving increased coding efficiency, better subjective picture quality, as well as an object-based video representation. MPEG-4 for instance, is the first standard allowing the transmission of arbitrarily shaped video objects (VOs).

According to the order of spatial and temporal features used, video object segmentation algorithms can be classified into three major categories: i) segmentation with spatial priority, ii) segmentation with temporal priority and iii) joint spatial and temporal segmentation [3]. In contrast to the first two categories 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. From a video object segmentation perspective, using a joint spatio-temporal strategy is superior to processing in either only the spatial or temporal domains, as it considers a video sequence as a spatio-temporal grouping of pixels. Existing spatio-temporal object segmentation techniques however, only consider pixel features, which tends to limit their performance in being able to segment arbitrary shaped objects.

Probabilistic space-time video object segmentation is one of the most popular spatio-temporal video segmentation techniques. It has a strong theoretical basis, with the segmentation being formulated within a statistical probabilistic framework. In [4], 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 comprising space, colour and time. The feature vectors are characterised by the GMM with parameter estimation achieved using the established *expectation maximization* (EM) algorithm [5]. A 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 some overlapping 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 applied, it has the fundamental drawback of being very dependent on the pixel features. In addition, the computational complexity increases as a direct result of considering BOF overlaps.

Colour and spatial location are very important features for object representation, though they are insufficient to represent all types of objects, as there are typically a huge number of objects and a myriad of variations amongst them. For this reason, in most cases, colour and spatial features alone fail to precisely approximate objects. This motivated consideration of integrating visual attributes into the process, so they intrinsically represent an

object, with the most important perceptual attribute of any object being shape as it provides valuable clues for humans in both distinguishing and recognising objects.

Ahmed et al [6] first proposed incorporating shape information into probabilistic spatio-temporal video object segmentation (PST-S) framework by using the original PST method in [4] and extending the elliptical shape used in video object tracking in [7]. The PST-S technique automatically extracts and incorporates shape information in the GMM model using a number of chord lengths passing through the centre of each object. The performance of the PST-S method was shown to be improved in comparison with the original PST technique [4], though one limitation was that since PST-S determines the probability of a pixel to be in a particular cluster, using the Gaussian distribution function always affords the highest probability to that cluster centre. This means the probability value decreases with distance from the centre, with the consequence that within a cluster the lowest probability occurs at the boundary of a shape, which is strongly correlated with the spatial features used already in the segmentation. This compromised the overall improvement in performance and motivated the incorporation of region-based shape information, where all pixels inside a particular shape will have similar probability.

This paper specifically addresses this problem by proposing a new probabilistic spatio-temporal video object segmentation technique incorporating region-based generic shape information (PST-RS) that employs the fundamental concepts of the uniform and confidence interval of the Gaussian distribution. The subjective evaluation of the proposed method has been compared with PST and PST-S using a number of standard video test sequences including *Salesman* and *Carphone*.

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

2. PROBABILISTIC SPATIO-TEMPORAL (PST) VIDEO OBJECT SEGMENTATION

In the PST algorithm, every pixel is represented by a six dimensional feature vector of space, colour and time. The *L,a,b* colour space is used to characterize the pixels as it is approximately uniform in perception and the distances in this space are meaningful [6].

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

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

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 using the *K*-means algorithm and iteratively obtains θ_{ML} from the following set of equations:

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

$$\hat{\alpha}_j \leftarrow \frac{1}{n}\sum_{i=1}^{n} p_{ij}, \quad \hat{\mu}_j \leftarrow \frac{\sum_{i=1}^{n} p_{ij} x_i}{\sum_{i=1}^{n} p_{ij}}$$

$$\hat{\Sigma}_j \leftarrow \frac{\sum_{i=1}^{n} p_{ij}\left(x_i-\hat{\mu}_j\right)\left(x_i-\hat{\mu}_j\right)^T}{\sum_{i=1}^{n} p_{ij}} \quad (4)$$

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

3. INCORPORATION OF REGION-BASED SHAPE INFORMATION

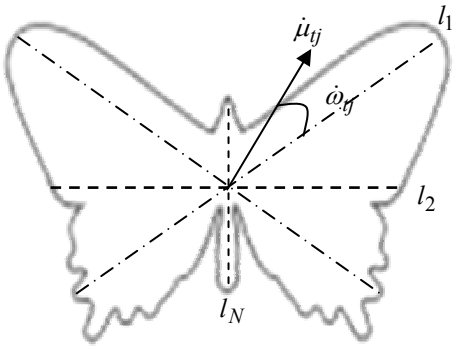


Fig.1. Illustration of shape using chords.

The video object tracking method employed for elliptical shape information in terms of the major and minor axes in a maximum *a posteriori* (MAP) framework in [7] is unable to consider arbitrary-shaped objects. The fundamental concept however, has been extended in PST-S [6] to facilitate the representation of generic shape information, using a strategy involving a series of chords passing through the centre of each object, which approximates the object shape as illustrated in Fig. 1.

The shape of each object is determined using the probabilistic segmentation algorithm [4] considering space and colour features described in Section 2. From the extracted shape contour, the major axis is determined and *N* chord lengths are generated at equal angular distances. The rotation and translation of an object is normalized by estimating the angle between the major axis of that object in the current and previous frames, as well as updating the object's centre. If the 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 [7] as:

$$O_{tj}(x_i) = \frac{1}{\sqrt{2\pi|\Sigma_{tj}|}} e^{-\frac{1}{2}(x_i-\mu_{tj})^T \Sigma_{tj}^{-1}(x_i-\mu_{tj})} \quad (5)$$

where μ_{tj} is the translation parameter. The covariance matrix Σ_{tj} is defined as:

$$\Sigma_{tj} = R^T(-\omega_{tj}) Diag\left[l_1^2,...,l_N^2\right] R(-\omega_{tj}) \quad (6)$$

where $l_1,...,l_N$ are *N* chord lengths as shown in Fig.1 and ω_{tj} is the rotation angle. The longest chord passing through the centre of the object is the major axis. Using the contour points on the object boundary, other chords are drawn through the centre at equal angular distances. As alluded in Section 1, the inherent limitation of this technique is that the prior function in (5) is directly correlated with spatial feature i.e. pixel location, as it gives the highest priority to the pixels belong to the centre of the shape and their priorities decrease with distance from the centre, which evidently limits the overall improvement of PST-S over PST.

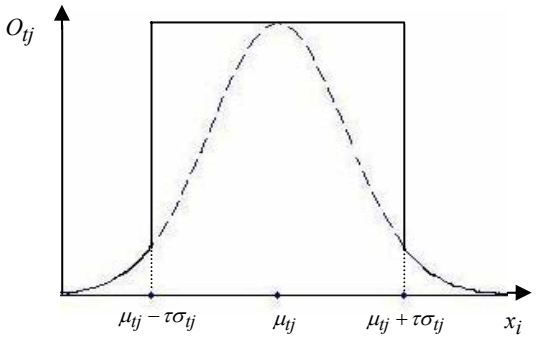


Fig. 2. Quasi-uniform pdf function.

To represent a shape using a region (silhouette) where all pixels inside a particular region have a similar probability, this paper proposes the following *quasi-uniform* probability density function utilising the basic concept of the confidence interval of the Gaussian distribution:

$$F_{tj}(x_i) = \begin{cases} O_{tj}(\mu_{tj}) & \text{for } \mu_{tj}-\tau\sigma_{tj} \le x_i \le \mu_{tj}+\tau\sigma_{tj} \\ O_{tj}(x_i) & \text{for all other } x_i \end{cases} \quad (7)$$

where σ_{tj} is the standard deviation of the j^{th} cluster and τ is the number of standard deviations which essentially represents the confidence interval, with for our proposed method $\tau = 2$ as shown by the solid lines in Fig. 2. The rationale for using this function is obvious from the statistical evidence [9] which shows the probability of a measurement from a Gaussian distribution falling within the confidence interval of *2* standard deviation ($2\sigma_{tj}$) of the mean μ_{tj} is 0.9544997. Hence, (7) ensures almost all the pixels within a shape boundary will have a similar priority so embodying the fundamental aim of a region-based approach for shape representation.

3.1. 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:

$$S_j(x_i|w_t) = w_t f_j(x_i) + (1-w_t) F_{tj}(x_i) \quad (8)$$

where $f_j(x_i)$ is the density function defined in (1) for the *j*-th object layer and w_t is the parameter that trades off between shape and spatio-color space, whose value is determined by maximizing the probability function in (8). The labelling (hard decision) of each pixel is chosen as the maximum *a posteriori* probability given by:

$$L(x_i) = \arg\max_j S_j(x_i) \quad (9)$$

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

$$P(L(x_i)=j) = S_j(x_i) / \sum_{j=1}^{k} S_j(x_i) \quad (10)$$

A key feature of the proposed PST-RS video segmentation technique is that the probabilistic incorporation of shape is represented based upon region (silhouette) information enclosed by a contour 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 block of frames (BOF) in order to find a correspondence between every pair of adjacent BOFs [4]. This reduces the amount of redundant information that has to be processed during object segmentation [6]. Algorithm 1 details all the steps involved in the PST-RS method.

Algorithm 1: *Probabilistic video segmentation using region-based shape information (PST-RS)*

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

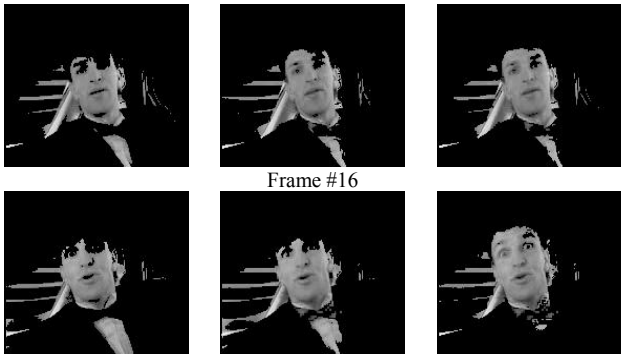
1. Extract feature vectors from a video frame and initialise GMM model parameters (1) using *K-means* algorithm.
2. Apply EM algorithm to estimate GMM model parameters using (3) and (4) in **Step 1**.
3. Select model using the MDL principle.
4. Determine object shape from the clusters and define shape of a particular cluster with a specified number of chords passing through its centre.
5. Find major axis and centre of the cluster.
6. Determine the quasi-uniform probability of pixels (7)
7. Calculate the joint probability of an object by (8).
8. Label each pixel using (10).
9. STOP.

4. SIMULATION RESULTS

The new video segmentation algorithm 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 true colour QCIF standard video test sequences of frame-size 176×144 pixels. Fig.3. shows the representative examples and their respective frame numbers for the *Carphone* and *Salesman* video test

(a) *Carphone* #16 (a) *Salesman* #16

Fig. 3. Original video frames

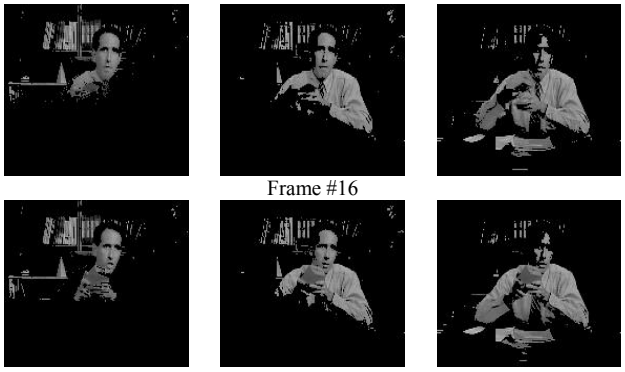


Frame #16

Frame #69

(a) PST (b) PST-S (c) PST-RS

Fig. 4. *Carphone* sequence



Frame #16

Frame #48

(a) PST (b) PST-S (c) PST-RS

Fig. 5. *Salesman* sequence

sequences, which has been widely used to evaluate video object segmentation performance [1], [6]. The *Salesman* sequence does not possess any global motion, but the motion of the non-rigid object (salesman) is significant in this sequence, especially in respect to the arm movements, while conversely the *Carphone* sequence possesses both global and object motion, due to the fast moving background visible through the car window and the object motion caused by passenger movement.

The respective segmentation results for particular frames from the two sequences produced using the original probabilistic space-time (PST) algorithm, PST-S and the new PST-RS segmentation techniques are shown in Fig. 4 and Fig. 5. If the results for frame #16 of *Carphone* in Figs. 4(a), 4(b) and 4(c) are compared with the original frame in Fig. 3(a), it is visually apparent, that despite the presence of global motion, many more pixels in both the forehead and chin regions have been accurately separated using the new shape-based PST-RS segmentation strategy. A similar observation can be made for the corresponding results of frame #16 of *Salesman* in Fig. 5 which reveals a considerable portion of the body has been correctly segmented and the background region reduced in Fig. 5(c) when compared with the results produced by PST and PST-S in Fig. 5(a) and 5(b), so vindicating the incorporation of region-based generic shape information into the segmentation framework. To confirm these results, two additional frame sets have been included for the respective test sequences (#69 and #48 for *Carphone* and *Salesman* respectively) in Fig. 4 and Fig. 5, which again endorse the judgement that integration of region-based shape information has improved the overall video object segmentation performance compared with the relevant contemporary techniques.

5. CONCLUSIONS

Automatic video object segmentation techniques mostly rely on the pixel features ignoring the shape of the object to be segmented hence these techniques do not perform well for all types of video objects. It is found from the literature that incorporation of shape information increases the performance of segmentation; however, the shape information directly correlated with spatial feature already used in the segmentation process theoretically limits its improvement. To address this limitation, this paper has introduced a new video object segmentation technique that seamlessly incorporates region-based 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 region-based generic shape information consistently provides in terms of improving the overall segmentation performance.

6. REFERENCES

[1] H. Musmann, M. H¨otter, and J. Ostermann, “Object-oriented analysis-synthesis coding of moving images,” Image Commun., vol. 1, pp. 117–132, Oct. 1989.

[2] M. Kunt, “Second-generation image coding techniques,” Proc. of the IEEE, vol. 73, pp. 549–574, Apr. 1985.

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

[4] 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.

[5] 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.

[6] R. Ahmed, G. C. Karmakar and L. S. Dooley, “Probabilistic Spatio-Temporal Video Object Segmentation Incorporating Shape Information,” Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. II, pp. 645-648, May 2006.

[7] 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.

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

[9] J. F. Kenney and E. S. Keeping, “Confidence Limits for the Binomial Parameter” and “Confidence Interval Charts,” Mathematics of Statistics, pt. 1, 3rd ed., pp. 167-169, Princeton, NJ: Van Nostrand, 1962.