PROBABILISTIC SPATIO-TEMPORAL VIDEO OBJECT SEGMENTATION USING *A PRIORI* SHAPE DESCRIPTOR

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

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

ABSTRACT

Since shape is regarded as one of the most important attributes of visualisation, it plays a pivotal role in semantic video object segmentation applications. One of the major objectives for the research community is to segment specific objects of interest from a video sequence using prescribed shape descriptors in a diverse range of applications from video surveillance and object tracking through to medical imaging. This paper addresses this challenge by presenting a new *probabilistic spatio-temporal* (PST) video object segmentation algorithm that incorporates *a priori* generic shape descriptor representations of particular objects in a sequence. The algorithm provides considerable improvement in perceptual picture quality compared with the existing PST segmentation technique, with the numerical analysis corroborating the superior subjective segmentation performance achieved.

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

1. INTRODUCTION

Semantic video object segmentation is one of the most important and demanding contemporary research topics embracing many disparate application domains with the main focus being, though by no means limited to, surveillance and object tracking, content-based video storage and retrieval, video footage analysis for various investigative purposes, traffic systems, video coding, editing, and medical diagnosis. A semantic video object can be defined as a collection of image pixels projecting a real object of interest in successive image planes of a video sequence. The semantics may change according to either the application or user-interest. The principal requirements of semantic video object segmentation are the precise definition of an object's boundary, i.e., spatial accuracy, and the subsequent maintenance of this accuracy throughout the video sequence, i.e., temporal coherence [1]. While humans are able to effortlessly differentiate video objects, fully automated computer-based object segmentation still remains a very challenging problem for the multimedia technology research community.

From a video object segmentation perspective, using a joint spatio-temporal strategy is superior to separately processing in either the spatial or temporal domains, as it considers a video sequence as a spatio-temporal grouping of pixels [2]. Since video objects of interests often exhibit certain characteristics like motion and/or spatial correlation, joint spatio-temporal techniques have the distinct advantage of embracing both spatial and temporal features with the segmentation framework, especially as psychologists have long recognised that the human visual system often finds salient structures jointly in space and time [3].

One of the most popular spatio-temporal video segmentation techniques is the *probabilistic space-time* (PST) approach which has a strong theoretical basis, with the segmentation task formulated in a statistical probabilistic framework [4]. The segmentation uses a piecewise Gaussian mixture model (GMM) to map a video sequence into a six-dimensional feature vector comprising space, colour and time. The feature vector is characterised by the GMM with parameter estimation achieved using the well-established *expectation maximization* (EM) algorithm [5]. A key attribute of this technique is that it analyses video frames as a single entity for model estimation purposes, so a contiguous *block of frames* (BOF) is considered and model estimation performed within each individual BOF under the assumption that all object motion is approximately linear.

While the basic PST concept has been widely applied [3][6], it has the fundamental limitation of being very dependent on pixel features. Colour and spatial location are key features for object representation, though they are insufficient to represent all object types as typical video sequences usually have a large number of perceptual objects and a multitude of variations amongst them. Even a single object can have multiple constituent objects comprised of many different colours. For this reason, colour and spatial features alone fail to approximate satisfactorily all objects, which motivated investigation of alternative visual attributes that more intrinsically represent objects. One of the most natural perceptual features of any object is its shape as this provides valuable cues for humans in being able to both recognise and distinguish it in any pictorial scene.

Ahmed et al. [6] seamlessly embedded shape information into the PST video object segmentation framework using the original PST method [4], though this possessed the crucial drawback of being dependent upon the initial segmentation, as the object shape was automatically extracted from the initial segmentation process. As alluded earlier, defining semantic video objects is an intractable task so that all the aforementioned approaches often fail to extract specific objects of interest. To our knowledge, there is currently no existing technique able to automatically segment a particular object from a video sequence based upon a prescribed arbitrary shape descriptor representation.

This paper presents an innovative PST-based strategy for segmenting particular video objects by incorporating *a priori* shape descriptor (PST-SD) representations within the PST video object segmentation framework. The two key distinguishing features of this new approach are that it reduces both the inherent dependency of earlier techniques on pixel-based features and its reliance on the initial segmentation phase. The performance of the PST-SD algorithm has been evaluated using different video test sequences to demonstrate its merits, with an objective evaluation technique [1] also being applied to validate the perceptual segmentation performance.

The remainder of the paper is organised as follows: In Section 2 the theoretical foundations of PST video object segmentation are outlined, while the theory in developing suitable shape descriptors and their subsequent integration into the PST framework is detailed in Section 3. A full analysis of the experimental results is presented in Section 4, with some concluding remarks given in Section 5.

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

In the original PST algorithm [4], 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:

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

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, where d is the feature vector dimension. The *maximum likelihood* (ML) estimation of θ for a set of feature vectors $x_1,...,x_n$ is given by:

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

The EM algorithm initialized using the well-established *K*-means algorithm, is then applied to iteratively estimate parameters θ_{ML} for the GMM [5] using the following set of equations:

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

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

with the information-theoretic framework based upon the principle of *minimum description length* (MDL) being employed to determine the model order i.e., the number of clusters *k* [2].

3. INCORPORATION OF SHAPE DESCRIPTORS INTO PST FRAMEWORK (PST-SD)

As the definition of semantic is highly dependent on interpretation, it is wholly unrealistic to segment semantic video objects without having any clues or *a priori* knowledge about the objects to be segmented. In the proposed method, the shape descriptor of a particular object of interest is provided *a priori* using a piecewise linear approximation of a shape contour derived from a set of significant points (SP) [8], as shown in the object shape example in Fig. 1 from a frame of the *Miss America* video test sequence.

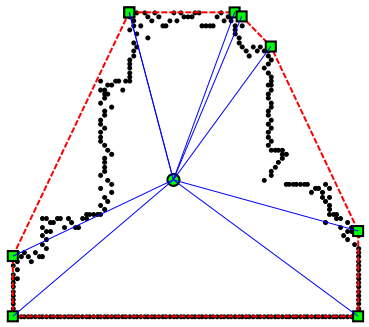

Figure 1: Significant point representation for the object (woman) shape of *Miss America* sequence.

The shape descriptor is made invariant to translation, scaling and rotation by using the *window-to-viewport transformation* [9] and then compared with the segmented regions produced by the GMM clustering strategy described in Section 2, for all intersecting points in order to find the best matching region. The region with the highest number of intersection points that is enclosed by a given shape descriptor is then selected as the video object of interest. The pre-generated polygon using the SP is then superimposed on the selected region with their respective centres aligned.

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 as:

$$O_{tj}\left(x_i\right)=\frac{1}{\sqrt{2\pi\left|\Sigma_t\right|}} e^{-\frac{1}{2}\left(x_i-\mu_t\right)^T\Sigma_t^{-1}\left(x_i-\mu_t\right)} \tag{5}$$

where μ_t is the centre of the selected object of interest and the covariance matrix Σ_t is determined using the centroidal distances from the SP shown in Fig. 1.

Incorporating region-based shape information is vital for precise segmentation of video objects [6]. To represent a particular shape with a region (silhouette) where all pixels enclosed by the region have similar probability, a *quasi-uniform* probability density function is employed. This embraces the notion of a uniform distribution with a confidence interval introduced as follows:

$$F_{tj}\left(x_i\right)=\begin{cases} O_{tj}\left(\mu_t\right) & \text{for } \mu_t-\tau\sigma_t \le x_i \le \mu_t+\tau\sigma_t \\ O_{tj}\left(x_i\right) & \text{for all other } x_i \end{cases} \tag{6}$$

where σ_t is the standard deviation and τ the number of standard deviations, which essentially defines the confidence interval set at $\tau=2$ in the PST-SD method. The rationale behind this function is that the probability of a measurement from a Gaussian distribution

falling within the confidence interval $2\sigma_t$ of the mean μ_t is 0.9544997, so (6) ensures almost all the pixels within a shape boundary have a similar priority so embodying the core aim of integrating the given shape descriptor into a region-based approach for shape representation.

Pixel Labelling: The labelling (hard decision) of each pixel is chosen as the *maximum a posteriori* probability given by:

$$L(x_i) = \arg\max_j F_{tj}(x_i) \quad (7)$$

while the confidence level (soft decision) of a pixel belonging to cluster j is defined as:

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

A key feature of the PST-SD algorithm is that the incorporation of a given shape descriptor is represented by region (silhouette) information which is enclosed by a prescribed contour. This influences significantly the probability of pixels to be either labelled or assigned to a particular cluster thereby more precisely representing an object shape. Algorithm 1 formalises all the various steps involved in the new video segmentation technique.

Algorithm 1: *Probabilistic video object segmentation using a priori shape information (PST-SD)*

Precondition: Video test sequence

Post condition: Segmented video object sequence.

1. Represent *a priori* shape information of the object to be segmented using the SP of the shape contour.
2. Extract feature vectors from the sequence to be segmented and initialise GMM model parameters (1) using *K-means* algorithm.
3. Apply EM algorithm to estimate GMM model parameters using (3) and (4) in **STEP 2**.
4. Select model using the MDL principle.
5. Determine object shape from segmented clusters.
6. Normalize each object shape and given shape descriptor to make them rotation, translation, and scale invariant.
7. Find the closest matching object with the given shape calculating intersection points by aligning shape centres.
8. Determine *quasi-uniform probability* of pixels by (6).
9. Label every pixel using (7).
10. STOP.

4. SIMULATION RESULTS

The PST-SD video segmentation algorithm has been implemented using MATLAB 7.2.0.232 (R2006a) running on Pentium-IV, 2.4 GHz CPU with 512 MB of memory. Experiments were conducted using true colour standard video test sequences of frame-size 96×72 pixels. Fig.2. shows three examples and their respective frame numbers for the popular *Miss America*, *Foreman*, and *Akiyo* sequences which have been widely used to evaluate video object segmentation performance [3][6]. The *Miss America* sequence does not possess any global motion, though the motion of the non-rigid object is significant, while conversely the *Foreman* sequence possesses both global as well as relatively high object motion. The respective results for particular frames from these three sequences for both the original PST [4] and new *a priori* shape-based object segmentation algorithm are shown in Figs. 3 and 4. The requisite shape contour information for the object of interest (person) in the *Miss America* sequence was provided *a priori*. It is readily evident from Fig 3b that incorporating the prescribed shape descriptor in

(a) *Miss America* #10 (b) *Foreman*# 10 (c) *Akiyo*# 10

Figure 2: Original video frames

Frame#10 *Frame*#54 *Frame*#89

(a) PST approach

(b) Proposed PST-SD approach

Figure 3: Results of segmentation for *Miss America* sequence

Frame#10 *Frame*#59 *Frame*#80

(a) PST approach

(b) Proposed PST-SD approach

Figure 4: Results of segmentation for *Foreman* sequence

Frame#10 *Frame*#35 *Frame*#80

(a) With PST approach

(b) Using the PST-SD approach with the *a priori* shape descriptor for *Miss America*

Figure 5: Results of segmentation for *Akiyo* sequence

the region-based strategy developed in Section 3, has enabled the precise segmentation of the object of interest including most notably the red jacket. This contrasts somewhat starkly with the corresponding results for the PST algorithm (Fig 3a). Indeed the results reveal that many misclassified pixels, especially in the vicinity of the chin and body have now been correctly segmented using the PST-SD algorithm. Similar observations can be made for the *Foreman* sequence, where the *a priori* shape descriptor for the

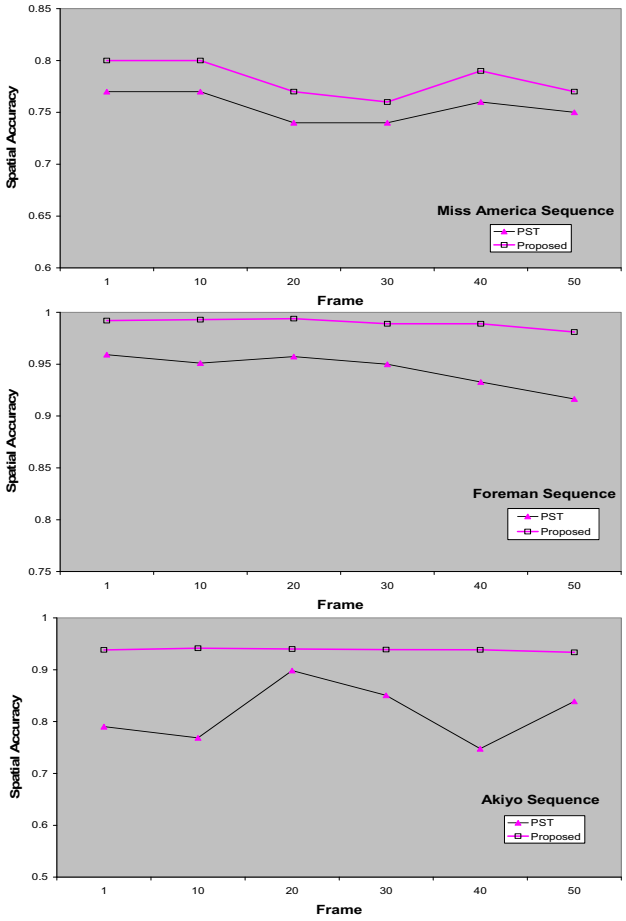

Figure 6: Objective performance evaluation results.

object of interest (man) was again provided. Representative frames from this more complex sequence are displayed in Fig. 4b, which confirm a striking improvement in the perceptual segmentation performance compared with for the original PST approach in Fig. 4a, with a conspicuously large number of misclassified pixels from the background now being correctly classified.

A further series of experiments were conducted to assess the PST-SD performance in handling *a priori* generic shape descriptors for specific classes of object such as human outlines or contours. The head and shoulder object (woman) from the popular *Akiyo* sequence was segmented using exactly the same shape descriptor as employed for *Miss America*. As anticipated PST-SD correctly segmented the object of interest with only a very small number of misclassified pixels in the hair region as evidenced in Fig. 5. This example illustrates the potential to derive improved segmentation performance using PST-SD with a series of generalised shape descriptor templates for different object classes.

To numerically confirm the perceptual findings upon the video object segmentation, a quantitative analysis [1] was undertaken for all three video sequences. The discrepancy metric is based on the objective error which measures spatial accuracy with respect to a reference segmentation using false positive and false negative errors respectively so:

$$err_{fp}(n) = card\left(test(n) \cap \overline{ref(n)}\right) \quad (9)$$

$$err_{fn}(n) = card\left(\overline{test(n)} \cap ref(n)\right) \quad (10)$$

where $test(n)$ is the set of pixels from a segmented object in frame n, $ref(n)$ is the reference segmentation and $card()$ represents the cardinality of a set. The spatial error of frame n is then given by:

$$err(n) = err_{fp}(n) + err_{fn}(n) \quad (11)$$

which is then normalized to obtain the measure of spatial accuracy:

$$S_{acc}(n) = 1 - err_{norm}(n) \quad (12)$$

Fig. 6 plots the comparative performance of the two techniques for the three test sequences in terms of spatial accuracy [1] which is concomitantly indicative of temporal coherence, thus endorsing the judgment that integrating *a priori* shape descriptor in the video object segmentation framework significantly improves its overall performance.

5. CONCLUSION

Semantic video object segmentation techniques have traditionally relied on pixel-based features and so do not integrate object shape information thereby hindering their generalisation capability for segmenting all video object types and user-specific requirements. A major goal of the relevant research community has been to be able to both segment and/or track a particular object shape of interest in a video sequence. This paper has introduced a new video object segmentation technique that seamlessly incorporates generic shape descriptor information about objects of interest into the probabilistic spatio-temporal framework. Both qualitative and quantitative empirical results for a number of disparate video test sequences have corroborated the efficacy of integrating *a priori* shape descriptor information into the framework by consistently providing both superior performance and achieving *semantics* in object segmentation.

6. REFERENCES

[1] A. Cavallaro, "From Visual information to Knowledge: Semantic Video Object Segmentation, Tracking and Description," PhD Thesis, University of Trieste, Italy, 2002.

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

[3] X. Song, and G. Fan, "Joint Key-Frame Extraction and Object Segmentation for Content-Based Video Analysis," IEEE Trans. on CSVT, 16(7), pp. 904-914, July 2006.

[4] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic Space-Time Video Modeling via Piecewise GMM," IEEE Trans. on PAMI, 26(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, 39(1), pp. 1-38, 1997.

[6] R. Ahmed, G. C. Karmakar and L. S. Dooley, "Region-Based Shape Incorporation for Probabilistic Spatio-Temporal Video Object Segmentation," IEEE Int. Conf. on Image Processing, Atlanta, USA, in press, October 2006.

[7] H. Tao, H. S. Sawhney, and R. Kumar, "Object Tracking with Bayesian Estimation of Dynamic Layer Representations", IEEE Trans. on PAMI, 24(1), pp. 75-89, January 2002.

[8] L. da F. Costa, and R. M. Cesar Jr., "Shape Analysis and Classification: Theory and Practice," pp 331-419, CRC Press LLC, 2001.

[9] J. D. Foley et al., "Computer Graphics: Principles and Practice," 2nd ed. In C, Addison-Wesley Pub. Co. Inc., 1999.