INCORPORATION OF TEXTURE INFORMATION FOR JOINT SPATIO-TEMPORAL PROBABILISTIC VIDEO OBJECT SEGMENTATION

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

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

ABSTRACT

Embedding textural information into the *probabilistic spatio-temporal* (PST) video object segmentation is very important for achieving better segmentation, since this is one of the key perceptual attributes of any object. Existing video segmentation techniques however, ignore this feature because of the underlying difficulty in defining and hence characterizing a texture, which theoretically limits their segmentation performance. To address this problem, this paper proposes a new video object segmentation algorithm that involves a strategy to seamlessly incorporate texture information as a pixel feature in the PST framework. Experimental results for a variety of standard test sequences reveal a significant performance improvement in the quality of the video object segmentation achieved in comparison with the original PST method.

Index Terms— Image sequence analysis, video segmentation, joint spatio-temporal, machine vision, image texture.

1. INTRODUCTION

Video object segmentation is of pivotal importance as it traverses many diverse 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. This is one of the most demanding and challenging contemporary research issues as fully automatic computer-based video object segmentation still remains a problematic objective for the multimedia research community.

Video object segmentation algorithms are usually classified into three major categories based on the order of spatial and temporal features used: 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 category 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 (PST) object segmentation is one of the most popular spatio-temporal techniques for video sequences. It has a strong theoretical basis, with the segmentation being formulated within a statistical probabilistic framework. In [2], a PST-based video segmentation approach using a piecewise Gaussian mixture model (GMM) has been proposed, which maps 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 [3]. 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 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 highly dependent on the pixel features. Colour and spatial location are important features for object representation, though they alone 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 many cases colour and spatial features fail to precisely approximate objects as they do not directly consider texture, which is one of the most vital, if paradoxically, ill-defined attributes of an image. Texture relates to the specific structure of either a visual or tactile surface characteristic of a particular object and defines the structure or composition of that object with regards to its components [4].

GMM-based PST segmentation algorithms segment objects based upon pixel level classifications which are common in video segmentation. In contrast, texture approximation techniques mainly calculate this feature for the entire image [5] and since texture represents surface characteristics of a particular object (image), it poses the interesting question as to *how texture can be defined for a particular pixel*. Inevitably it is thus very challenging to incorporate texture information into the PST framework, and while the importance of embedding texture information has already been clearly articulated [2], to the best of our knowledge, it has yet to be successfully achieved.

This paper presents a new PST-based video object segmentation algorithm that involves the innovative idea of incorporating texture information as a pixel feature into the PST foundation [2] using the standard deviation of luminance values of a group of neighbouring pixels. It also addresses the ubiquitous limitation of computational complexity for GMM based techniques by selecting a set of key-frames from the video sequence and then applying the segmentation algorithm to only the key-frames in order to achieve greater efficiency in the segmentation process. The new method has been evaluated and compared with the original PST technique [2] using a number of test video sequences including *Silent, Mother and Daughter,* and *Akiyo* to fully demonstrate its meritorious performance. The computational time complexity has also been compared with [2], which represents the efficiency of the proposed algorithm in terms of CPU time.

The remainder of this paper is as follows: In Section 2 the theoretical foundations of the PST object segmentation technique are briefly outlined, while the fundamental theory in representing texture 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. PST VIDEO SEGMENTATION FRAMEWORK

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 [3] 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,

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

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

The information-theoretic framework based upon the principle of Minimum description length (MDL) [6] is employed for model selection, i.e., selecting the most appropriate number of clusters.

3. INCORPORATION OF TEXTURE INFORMATION

As alluded in Section 1, texture is one of the most important attributes of a video sequence representing the structural arrangements of the surface as well as the relations amongst them. There are numerous qualitative structural features including fineness, coarseness, lineation, smoothness, granulation, directionality, roughness, regularity and randomness, which can informally assist in providing a suitable definition of texture [4], though while these particular features help to discriminate the desired texture types by defining a spatial arrangement of specific texture constituents, this heuristic classification are too fuzzy to form a basis for formal definitions and build computational models of image textures. Such models to describe natural images however, are essential in order to analyse and synthesise textures. A computational model of a texture involves a particular set of basic constituents, or elements

(a) $N=4$ (b) $N=8$

Fig. 1. Possible pixel neighbourhood configurations

and rules of their spatial arrangement, and in terms of texture elements, a homogeneous textured image frame has the basic property of spatial self similarity supported by the geometrically similar pixel combination over the frame. While it is a very intractable problem to establish a generic formal definition for such self similarity or repetitiveness, this paper develops a novel approach for defining the texture feature [8] for the *i-th* candidate pixel (as illustrated by ⊗ in Fig.1) in a frame comprising *n* pixels, by considering the standard deviation of the luminance values of neighbouring pixels as follows,

$$\tau_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N+1}\left(L_{x_j} - \bar{L}\right)^2}, \qquad i = 1,2,...,n \tag{5}$$

where N is the number of neighbouring pixels (represented by ⊕ in Fig.1) and L_{x_j} is the luminance value of a pixel x_j. τ_i is then appended to the feature vector described in Section 2 as a further dimension to apply for parameter estimation using GMM as depicted in (1)-(4).

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

$$Lab(x_i) = \arg\max_j f_j(x_i) \tag{6}$$

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

$$P(Lab(x_i) = j) = f_j(x_i) / \sum_{j=1}^{k} f_j(x_i) \tag{7}$$

Algorithm 1: *Probabilistic video object segmentation incorporating texture information*

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

1. Select key-frames from the video sequence.
2. Extract space and colour features of each pixel of a video frame.
3. Calculate the standard deviation of luminance values of a group of neighbouring pixels using (5) and append the values in the feature vector extracted in **Step 2.**
4. Initialise GMM model parameters (1) using *K-means* algorithm.
5. Apply EM algorithm to estimate GMM model parameters using (3) and (4).
6. Select model using the MDL principle.
7. Label each pixel using (6).
8. STOP.

3.1. Computational Time Complexity

The GMM has to handle a large dimensional feature vector of space, colour, time, and texture. A measure has therefore been taken, to reduce the computational time complexity by adopting a strategy of selecting key-frames *a priori* from the video sequence to be segmented by using the method in [9]. The proposed algorithm is then employed on the video sequence comprised of the selected key-frames only. This strategy is shown to dramatically reduce the overall time complexity, as is evidenced in Table 1.

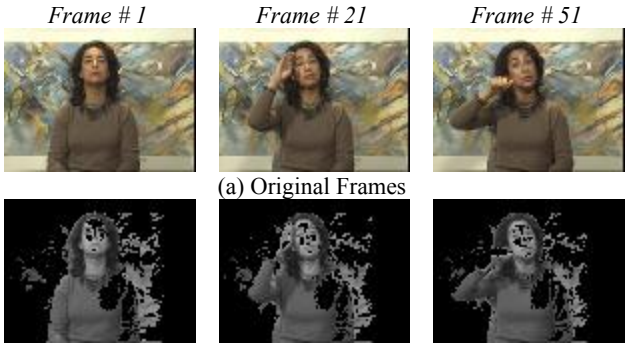
A key feature of the proposed video segmentation technique is that image texture information is represented as a pixel feature and the probabilistic incorporation of texture impacts on the probability of pixels to be either labelled or assigned to a particular cluster. Algorithm 1 details all the various steps involved in the new proposed method.

4. SIMULATION RESULTS

The new 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 and taking $N = 4$ in (5). Figs.2-4 show the representative examples and their respective frame numbers for the *Silent*, *Mother and Daughter*, and *Akiyo* video test sequences which has been widely used to evaluate video object segmentation performances. The *Silent* sequence does not possess any global motion, but the motion of the non-rigid object (the lady) is significant, especially in respect to the arm movements, with the background being both complex and highly textured. Conversely the *Mother and Daughter* sequence possesses object motion with the presence of multiple objects, while the *Akiyo* sequence has low object motion in the foreground but a complex background.

The respective segmentation results for randomly selected frames from the selected key-frames of the three test sequences produced using the original PST algorithm and the proposed segmentation techniques are shown in Figs. 2, 3, and 4. Visualization of Fig. 2b apparently presents the fact that there are many misclassified pixels throughout the body of the object and also in the background. As this particular video sequence has a very high textured background the incorporation of texture feature clearly gives better results which are evident in Fig. 2c. Comparison between Fig. 3b and 3c for *Mother and Daughter* sequence also suggests that a huge number of pixels have been correctly classified with the new approach, with similar observations being clearly evident in the forehead and hair region of *Akiyo* sequence in Fig.4c, which also depicts the significance of the proposed approach in terms of perceptual quality.

To corroborate the improvement in computational time complexity in terms of CPU time by adopting the strategy of key-frame selection discussed in Section 4, Table 1 displays the average CPU times required to segment the 100 frame test sequences. The number of key-frames selected for a 100 frame sequence of *Silent* was 11 while for *Mother and Daughter* and *Akiyo* they were 11 and 7 respectively, so around a 90% reduction in CPU time has been achieved by adopting the key-frame selection strategy.

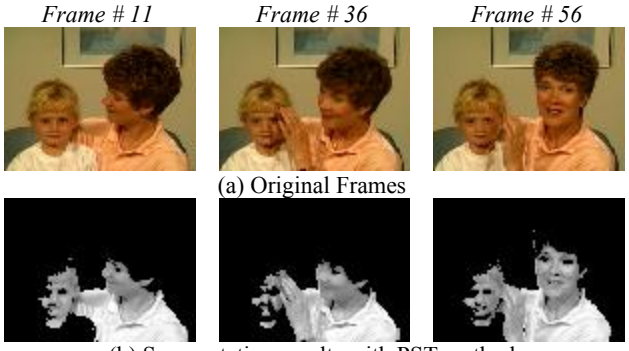


(a) Original Frames

(b) Segmentation results with PST method

(c) Segmentation results with proposed method

Fig. 2. *Silent* sequence

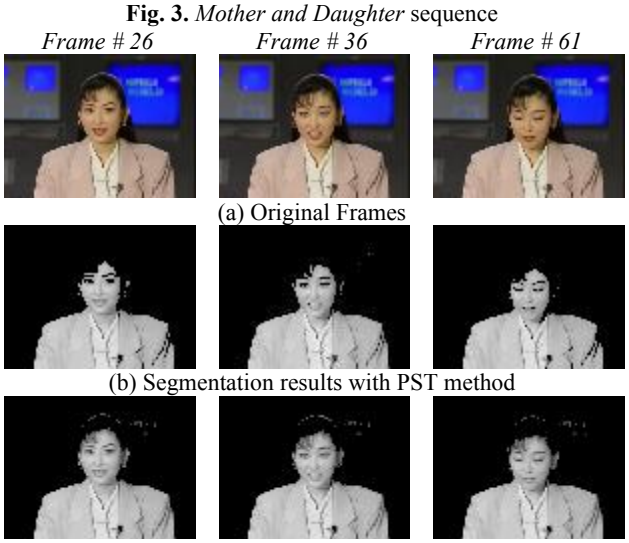


(a) Original Frames

(b) Segmentation results with PST method

(c) Segmentation results with proposed method

Fig. 3. *Mother and Daughter* sequence



(a) Original Frames

(b) Segmentation results with PST method

(c) Segmentation results with proposed method

Fig. 4. *Akiyo* sequence

Table 1: CPU time in seconds for a 100 frames video sequence

Video Sequence	PST method	Proposed method
Silent	169	19
Mother and Daughter	160	18
Akiyo	156	11

The CPU time required for selecting the key-frames was around 6 seconds for a 100 frames sequence.

5. CONCLUSION

Automatic video object segmentation techniques classify pixels of a frame based on their features to segment an object, and so do not perform well for all sequence types. Although defining image texture as a pixel feature is a challenging task, this paper has introduced an innovative strategy to incorporate texture information for each pixel in a frame and seamlessly integrate it into the *probabilistic spatio-temporal* segmentation framework. Experimental results upon a number of different test sequences have illustrated both the efficacy and benefit that incorporating texture information consistently bestows in terms of enhancing the overall video segmentation performance.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] G. L. Gimel'farb, "Image Textures and Gibbs Random Fields," Kluwer Academic Publishers, 1999.

[5] G. C. Karmakar, "An Integrated Fuzzy Rule-Based Image Segmentation Framework," PhD Thesis, Monash University, Australia, 2002.

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

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

[8] R. C Gonzalez, R. E. Woods, and S. L. Eddins, "Digital Image Processing Using MATLAB," Pearson Prentice Hall, 2004.

[9] J. Rong, W. Jin, and L Wu, "Key Frame Extraction Using Inter-Shot Information," Proc. of the IEEE International Conference on Multimedia and Expo, pp. 571-574, June 2004.