Texture as a pixel feature for video object segmentation

Rakib Ahmed[1], *Gour C. Karmakar*[1] and *Laurence S. Dooley*[2]
[1]Gippsland School of Information Technology, Monash University, Australia
[2]Faculty of Maths, Computing and Technology, The Open University, United Kingdom

As texture represents one of the key perceptual attributes of any object, integrating textural information into existing video object segmentation frameworks affords the potential to achieve semantically improved performance. While object segmentation is fundamentally pixel-based classification, texture is normally defined for the entire image, which raises the question of how best to directly specify and characterise texture as a pixel feature. This letter introduces a generic strategy for representing textural information so it can be seamlessly incorporated as a pixel feature into any video object segmentation paradigm. Both numerical and perceptual results upon various test sequences reveal a considerable improvement in the object segmentation performance when textural information is embedded.

Introduction: Recent developments have principally focused upon automatic object segmentation strategies [1-4] because of their diverse application base. A statistical framework based on graphical model is found in [1], while an alternative approach [2, 3] employs a *Gaussian mixture model* (GMM) to represent spatio-temporal features. In [2], spatio-temporal 3D volume of pixels comprising the video is modelled by a GMM, while the strategy in [3] separates objects frame by frame.

However, fully automated object segmentation based on conventional low level homogeneity criteria is still in its infancy since an object typically comprises multiple colours, and/or non-homogeneous motion in different parts. Most natural surfaces exhibit texture as it is a primitive perceptual cue, which can be visually sensed, so providing the *feel* of an object. To successfully segment an object thus requires consideration to be given to texture, alongside other low level features, with all features having equal priority. Video object segmentation techniques typically separate objects based upon pixel-level classifications in contrast to popular textural approximation techniques normally calculated for either the entire image or a region. Recent developments such as *scale invariant feature transform* and *gradient location and orientation histogram* [5] have exhibited promising performance in describing region, though pragmatically it is infeasible to consider them with other pixel features as they are histograms generated from the region. Markov random field model [6] considers texture involving the interactions among neighbouring pixels assuming homogeneous pixel intensity and thus is unable to represent a real object. This provided the motivation to formulate a textural representation strategy that quantifies texture as a pixel feature and enables the information to be incorporated into any video object segmentation process alongside other low level features, to accomplish perceptually and semantically improved performance. The *standard deviation* (SD) and *fractal dimensions* (FD) of a group of neighbouring pixels have been separately considered to approximate texture as a pixel feature, with the potential of these two strategies being individually tested upon both a GMM-based *joint spatio-temporal* (JST) [2] and *context-based spatio-temporal* (CST) [3] video segmentation paradigm, to evince the generic nature of the solution using a number of standard test videos, with an objective evaluation metric [7] being introduced to quantitatively verify the perceptual findings.

Object segmentation using textural information: As texture is a contextual property, its definition must involve pixel intensity values in any spatial neighbourhood of an image frame. Two effective strategies are presented for defining pixel textural information in terms of its neighbourhood. The SD of pixel intensity values of neighbouring elements is one feasible descriptor of texture, such that the textural feature τ_i for pixel x_i in a frame comprising n pixels can be represented by the SD of the luminance of its neighbouring pixels [4]:

$$\tau_i = \sqrt{\frac{1}{H}\sum_{j=1}^{H+1}\left(Lu_{x_i} - \overline{Lu}\right)^2}, \qquad i = 1,2,...,n \qquad (1)$$

where H is the number of neighbouring pixels, Lu_{x_i} is the luminance of pixel x_i and $\overline{Lu}$ is the average luminance of H neighbouring pixels.

A second strategy using the well-accepted FD has been adopted to represent texture as a pixel feature. A bounded set S in a Euclidean η-space is self-similar whenever S is the union of C distinct copies of itself, each of which has been scaled down by a ratio r. The FD then provides a measure of surface roughness as:

$$F_D = log\, C / log(1/r) \qquad (2)$$

where the larger the value of F_D, the rougher the surface. While this can effectively represent surface texture, the major obstacle remains of how to incorporate this information as a pixel-based feature within a video segmentation model, as FD applies to the entire image. To address this limitation, the popular *differential box counting* (DBC) [8] method for FD is used to calculate the textural feature for a candidate pixel $x_{i,j}$ by introducing a *sliding window* (SW) of size $h \times h$ pixels instead of using the entire image [8]. Since the candidate pixel lies inside the SW, the calculated FD feature in effect represents the surface variations of its neighbouring pixels, and so can be used as the textural feature for the candidate pixel.

To determine the FD for a SW of size $h \times h$ pixels, let the scale down ratio be $r = \chi/h$, where the image grid size is $\chi \times \chi$ and a third coordinate is introduced to represent the intensity level of each grid comprising a column of boxes of size $\chi \times \chi \times \chi'$, which for 8-bit luminance samples implies $256/\chi' = h/\chi$. If the maximum and minimum intensity levels in the grid $G_{u,v}$ reside in boxes B_{max} and B_{min} respectively, then the surface variation represented by the thickness of the blanket covering the image surface on the grid is:

$$sv_{u,v} = B_{max} - B_{min} + 1 \quad (3)$$

while the contribution from all grids defining the blanket is:

$$C = \sum_{u,v} sv_{u,v} \quad (4)$$

The FD of a SW calculated using (2) and (4), then represents the textural feature for pixel $x_{i,j}$. The greater the number of grids, the finer the measure of surface roughness. Lengthening the window commensurately increases the computational time, as more grids are included in the DBC calculations, though the overall order of complexity remains unchanged.

To validate the efficacy of the approximated textural information in terms of video segmentation performance, the SD and FD representations, denoted respectively by SDT and FDT, have been correspondingly integrated within the JST [2] and CST [3] together with other low level pixel features, adding an extra dimension to the original feature vector set for the GMM.

Results: Simulations have been performed using MATLAB with true colour standard test sequences of frame-size 96×72. Fig 1 displays sample frames for the widely used *Table Tennis* (TT) and the highly complex colour ultrasound *Baby Beatrix* (BB) sequence of a moving foetus in a mother's womb which is especially significant for its potential applications in medical imaging.

The comparative results for the original JST [2] and CST [3] models and their respective SDT and FDT based paradigms are shown in Fig 1. Results for the TT sequence (Figs 1*b*-*d*) consistently confirm a considerable number of pixels been correctly classified by both JST-SDT and JST-FDT, especially in the vicinity of the tennis bat and hand, with both methods correctly extracting the whole body, in contrast to the original JST approach. As the background of this particular sequence has a complex texture, the inclusion of a pixel texture feature has enabled relevant pixels to be grouped into a separate cluster, and while the misclassification by JST along the table edge is not fully eradicated in JST-SDT, it has been successfully corrected by JST-FDT due its latent capacity to more effectively represent surface variation. Comparative analysis for CST model shown in Figs 1*e*-*g* also reveals a significant improvement in object segmentation with the addition of textural features in both CST-SDT and CST-FDT.

Segmentation results for the BB sequence in Figs 1*c-d* confirm that both JST-SDT and JST-FDT have corrected a significant number of misclassifications of the original JST result along the foetus region and also in background pixel classification for all the representative frames. The comparative results for the CST based methods in Fig 1*e*-g display a similar trend with the foetus being correctly segmented, concomitant again with a noticeably lower background pixel misclassification for FDT.

To numerically substantiate the perceptual results, a discrepancy metric [7] based upon the spatial accuracy using false positive and false negative errors with respect to the ground truth, is presented in Fig 2. It can be readily concluded for the TT sequence that by embedding textural information, both SDT and FDT consistently outperformed the original JST and CST models, in terms of temporal coherence. This shows that both JST and CST display greater robustness to motion variations when textural information is incorporated. A similar observation can be made with respect to the BB sequence corroborating that the inclusion of pixel-based texture improves the quality of object segmentation.

Conclusion: Texture is an important perceptual attribute of any object, so this letter has introduced an innovative strategy to represent textural information as a pixel feature and seamlessly integrate it into two popular and contemporary *spatio-temporal* segmentation frameworks. Improved video object segmentation has been achieved in both segmentation paradigms, with the theory developed being generic in that it affords an efficacious way of seamlessly integrating texture to improve the quality in any segmentation model that exploits low-level pixel features.

References

[1] Y. Wang, K.-F. Loe, T. Tan, and J.-K. Wu, "Spatiotemporal video segmentation based on graphical models," *IEEE Transactions on Image Processing,* vol. 14, pp. 937-947, July 2005.

[2] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, pp. 384-396, March 2004.

[3] J. Goldberger and H. Greenspan, "Context-based segmentation of image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 463-468, 2006.

[4] R. Ahmed, G. C. Karmakar, and L. S. Dooley, "Incorporation of texture information in the joint spatio-temporal probabilistic video object segmentation," in *IEEE International Conference on Image Processing (ICIP2007),* Texas, USA, 2007, pp. 293-296.

[5] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 1615-1630, 2005.

[6] H.-D. Li, M. Kallergi, L. P. Clarke, W. Qian, and R. A. Clark, "Markov random field model for mammogram segmentation," in *15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 1993, pp. 54 - 55.

[7] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Tracking video objects in cluttered background," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 15, pp. 575-584, 2005.

[8] G. C. Karmakar, L. S. Dooley, and M. Murshed, "Fuzzy rule for image segmentation incorporating texture features," in *IEEE International Conference on Image Processing (ICIP 2002),* New York, USA, 2002.

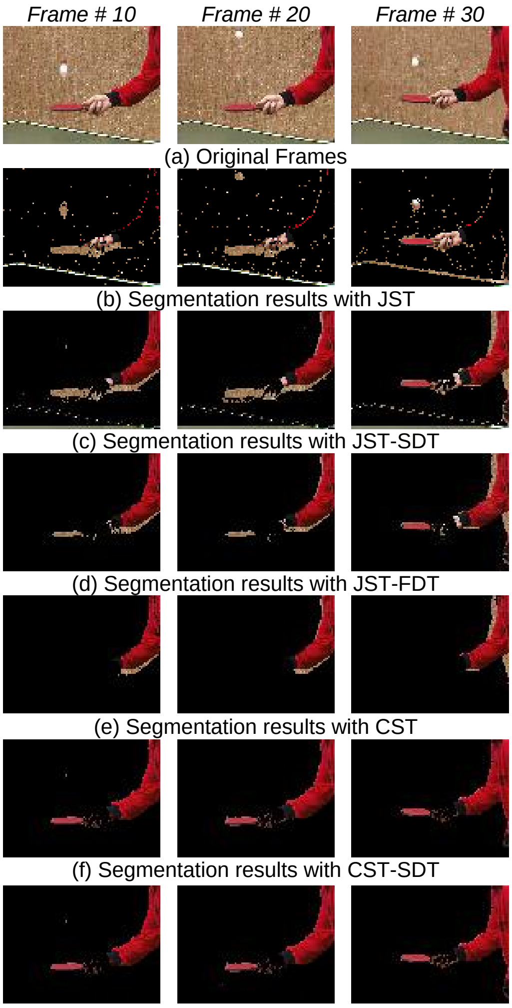
Figure Captions:

Fig 1: Segmentation results for TT sequence

Fig 2: Segmentation results for BB sequence

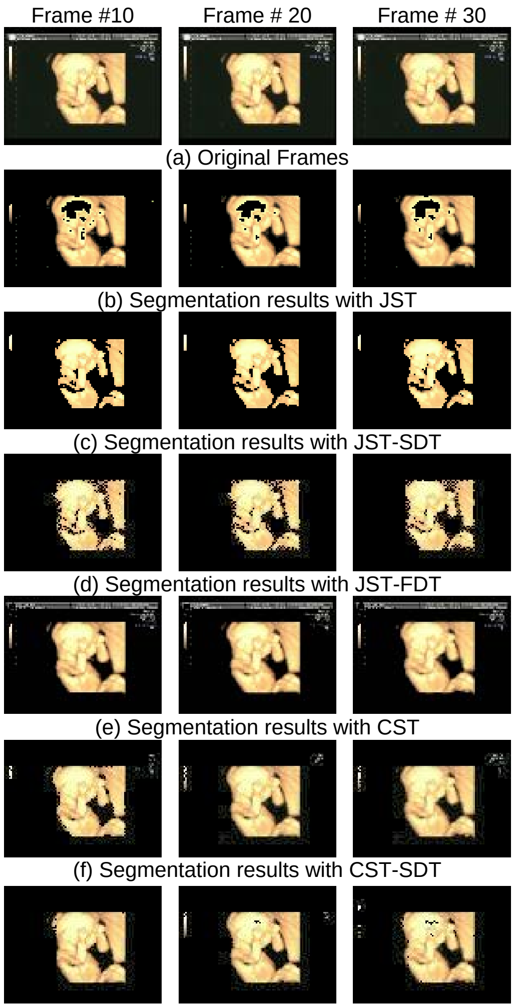
Fig 3: Spatial Accuracy

Figure 1



(g) Segmentation results with CST-FDT

Figure 2



(a) Original Frames

(b) Segmentation results with JST

(c) Segmentation results with JST-SDT

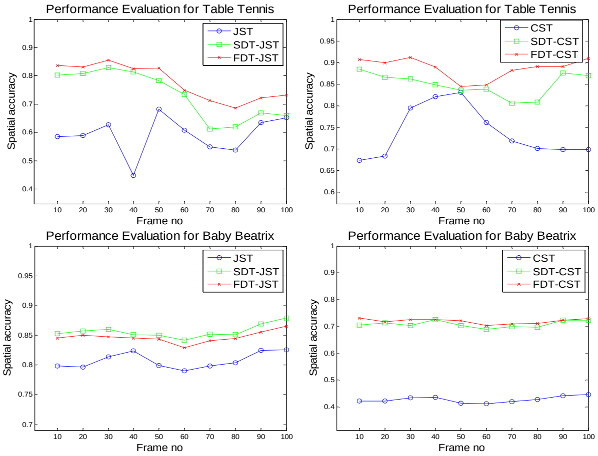
(d) Segmentation results with JST-FDT

(e) Segmentation results with CST

(f) Segmentation results with CST-SDT

(g) Segmentation results with CST-FDT

Figure 3



Performance Evaluation for Table Tennis
JST
SDT-JST
FDT-JST
Spatial accuracy
Frame no
Performance Evaluation for Table Tennis
CST
SDT-CST
FDT-CST
Spatial accuracy
Frame no
Performance Evaluation for Baby Beatrix
JST
SDT-JST
FDT-JST
Spatial accuracy
Frame no
Performance Evaluation for Baby Beatrix
CST
SDT-CST
FDT-CST
Spatial accuracy
Frame no