Region Growing for Motion Segmentation using an Attention Based Algorithm

Shijie Zhang and Fred Stentiford

Department of Electronic and Electrical Engineering

University College London, Adastral Park Campus

Abstract: This paper proposes a novel two-stage method for motion segmentation under stationary background conditions. First motion vectors are extracted using an attention based method. Then a region growing technique is applied to these vectors to obtain motion segmentation and completes motion segmentation with region matching. The algorithm is tested on various video data and experimental results show that the proposed approach can extract detailed motion information from non-rigid objects such as moving people.

1. Introduction

Motion segmentation is a process of segmenting foreground moving objects from the background scene in video sequences. It is a basic task for many computer vision applications, such as content-based video retrieval and indexing [1-3], object tracking [4], and object-based analysis and video coding [5,6].

Motion segmentation can be performed by either first estimating a field of motion parameters then segmenting it, or by applying jointly motion estimation and segmentation. In one class of approaches, motion parameters are computed and then segmented sequentially [4,6,7], while in the other class of methods motion estimation and segmentation are performed jointly. The problem in this class is usually formulated using the energy minimization or maximum a posteriori (MAP) probability frameworks [1,5,8-11] and solved by iterating between estimation and segmentation steps derived from the formulation.

Motion segmentation requires accurate estimation of movement. We adopt an attention based approach [12,13] to extract motion information which is then used in a region growing and matching process.

2. Motion segmentation framework outline

The proposed framework contains two stages. The first stage uses an attention based method [12,13] to estimate and extract motion information. The second stage then applies a region growing technique to motion vectors extracted. It completes motion segmentation with region matching. The outline of the algorithm is given below.

2.1 Motion estimation

Regions of static saliency have been identified using an attention method described in [14]. Those regions which are largely different to most of the other parts of the image will be salient and are likely to be in the foreground. This concept has been extended into the time domain and is applied to frames from video sequences to detect salient motion. The approach [12,13] does not require an initial segmentation process and depends only upon the detection of anomalous movements. The method estimates the shift of locations between frames by obtaining the distribution of displacements of corresponding salient features around these locations.

In this paper candidate regions of motion are detected by generating the intensity difference between the current frame and a background reference frame obtained by averaging a series of frames in an unchanging video sequence. A threshold is then applied producing a *potential motion template*. The intensity difference I_x between pixels *x* in the current frame and the reference is given by

$$I_{\boldsymbol{x}} = \{|r_2 - r_1| + |g_2 - g_1| + |b_2 - b_1|\}, \qquad (1)$$

where parameters (r_1, g_1, b_1) & (r_2, g_2, b_2) represent the rgb colour values for pixel $\boldsymbol{x}$ in reference frame and the current frame. The intensity I_x is calculated by taking the sum of the differences of rgb values between the two frames. The candidate regions *C* in the current frame are then identified where $I_x > T$. *T* is a threshold determined by an analysis of the image.

Let a pixel $\boldsymbol{x} = (x, y)$ in R_t correspond to colour components $\boldsymbol{a} = (r, g, b)$. Let $\boldsymbol{F(x)} = \boldsymbol{a}$ and let $\boldsymbol{x_0}$ be in frame R_t *at time* t. Consider a neighbourhood *G* of $\boldsymbol{x_0}$ within a window of radius ε where

$$\{\boldsymbol{x}_i' \in G \quad \textit{iff} \quad |x_0 - x'| \le \varepsilon\}. \qquad (2)$$

Select a set of *m* random points S_x in *G* (called a fork) where

$$S_x = \{\boldsymbol{x_1'}, \boldsymbol{x_2'}, ..., \boldsymbol{x_m'}\}. \qquad (3)$$

Forks are only generated which contain pixels that mismatch each other. This means that forks will be selected in image regions possessing high or certainly non-zero attention scores, such as on edges or other salient features. In this case the criteria is set so that at least one pixel in the fork will differ with one or more of the other fork pixels by more than δ in one or more of its rgb values i.e.

$$\left|F_k(\boldsymbol{x}_i') - F_k(\boldsymbol{x}_j')\right| > \delta_k, \quad \textit{for some } i, j, k. \qquad (4)$$

Define the radius of the region within which fork comparisons will be made as V (*view radius*). Randomly select another location $\boldsymbol{y_0}$ in the next frame R_{t+1} within a radius V of $\boldsymbol{x_0}$.

Define the second fork

$$S_y = \left\{\boldsymbol{y_1'}, \boldsymbol{y_2'}, ..., \boldsymbol{y_m'}\right\} \text{ where } \boldsymbol{x_0} - \boldsymbol{x_i'} = \boldsymbol{y_0} - \boldsymbol{y_i'} \text{ and } \left|y_0 - \boldsymbol{x_0}\right| \le V. \qquad (5)$$

S_y is a translated version of S_x. The fork centred on $\boldsymbol{x_0}$ is said to match that at $\boldsymbol{y_0}$ (S_x matches S_y) if all the colour components of corresponding pixels are within a threshold δ_k,

$$\left|F_k(\boldsymbol{x_i'}) - F_k(\boldsymbol{y_i'})\right| \le \delta_k, \quad k = r, g, b, \; i = 1,2,...,m. \qquad (6)$$

N attempts are made to find matches and the corresponding displacements are recorded as follows:

For the jth of $N_1 < N$ matches define the corresponding displacement between $\boldsymbol{x_0}$ and $\boldsymbol{y_0}$ as $\boldsymbol{\sigma_j^{t+1}} = (\sigma_p, \sigma_q)$ where

$$\sigma_p = \left|x_{0p} - y_{0p}\right|, \quad \sigma_q = \left|x_{0q} - y_{0q}\right|, \qquad (7)$$

and the cumulative displacements Δ and match counts Γ as

$$\left.\begin{aligned} \Delta(\boldsymbol{x_0}) &= \Delta(\boldsymbol{x_0}) + \boldsymbol{\sigma_j^{t+1}} \\ \Gamma(\boldsymbol{x_0}) &= \Gamma(\boldsymbol{x_0}) + 1 \end{aligned}\right\} \; j = 1,..., N_1 < N, \qquad (8)$$

where N_1 is the total number of matching forks and N is the total number of matching attempts.

The displacement $\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1}$ corresponding to pixel $\boldsymbol{x_0}$ averaged over the matching forks is

$$\overline{\boldsymbol{\sigma}}_{\boldsymbol{x_0}}^{t+1} = \frac{\Delta(\boldsymbol{x}_0)}{\Gamma(\boldsymbol{x}_0)}. \qquad (9)$$

This process is carried out for every pixel $\boldsymbol{x_0}$ in the candidate motion region R_t and M attempts are made to find an internally mismatching fork S_x. The displacements are saved in the motion vector map O_{MV} and O_{MV}'.

2.2 Motion segmentation

The motion vectors generated in the previous section tend to be associated with salient regions such as leading and trailing edges of moving objects; non-salient homogeneous regions are not assigned motion vectors and for this reason in the second stage a region growing algorithm is introduced which infers motion in these homogeneous regions. First homogeneous regions are identified. Then the position of the largest motion vector is taken as a seed for region growing and the value of this vector is assigned to pixels in the homogeneous region if this translation would lead to a pixel match in the next frame. This is repeated for the same homogeneous region to allow a different motion vector to be assigned to the remaining part of the same homogeneous region to obtain a match with the next frame. Regions which are changing shape would be affected by this process.

1. The location of the largest motion vector ϕ_1^i in O_{MV} is labelled as a starting pixel for region growing for region P_i (i=1 initially).
2. Its 8*p neighbourhood pixels (p=1 initially) are compared with the starting pixel for a colour match (equation (4)) and included in region P_i if a match is found.
3. Step 2 is repeated with p=p+1. The 8p-neighbourhood pixels are each included in region P_i if they match the starting pixel and are adjacent to a pixel already in region P_i.
4. Growing stops when no further pixels are included in region P_i in step 3.
5. Apply thresholding mask C to remove pixels grown beyond the candidate moving regions.
6. Assign ϕ_1^i to all pixels in region P_i if they match corresponding pixels in the next frame in colour. Update O_{MV}' with new assignments.
7. Search for 2nd largest motion vector ϕ_2^i from O_{MV}' in region P_i.

8. Assign ϕ_2^i to all homogeneous pixels in region P_i whose motion vectors are not already assigned with ϕ_1^i if they match corresponding pixels in the next frame.
9. Update $O_{MV}^{'}$ and O_{MV} with new assignments.
10. Repeat Step 1-9 for i = 1, …, Φ regions until growing ceases.

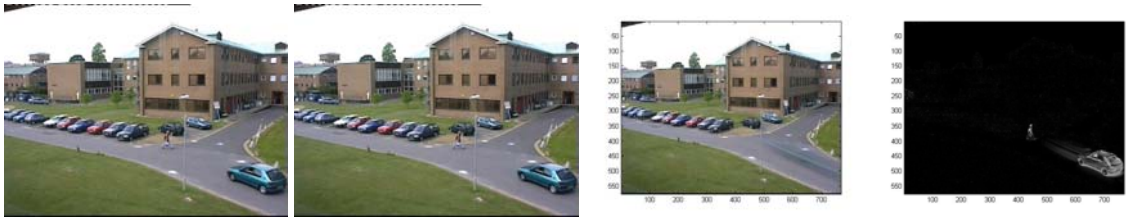
Seed motion vectors ϕ_1^i are rejected if their locations are not present in a difference frame between the current and next frame. This eliminates the spurious analysis of stationary objects not present in the reference frame.

3. Results and Discussion

The attention based region growing algorithm is illustrated on various data [15] both indoors and outdoors. The parameters of all experiments were M = 100, N = 10000, $\varepsilon = 1$, m = 2, $\boldsymbol{\delta}$ = (40,40,40), T = 90. V is selected according to the maximum velocity expected in the clip. Values of Φ in the results below reflect the point at which no further motion vectors are assigned.

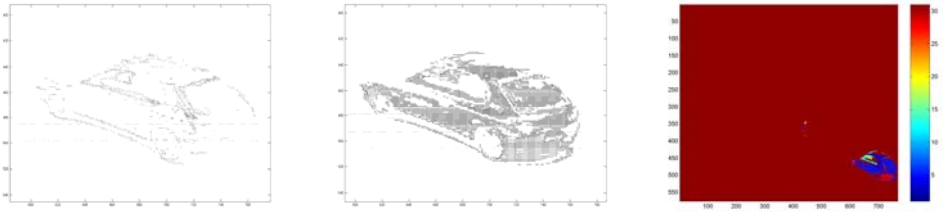
3.1 Reading University

A pair of 768x576 frames from a campus sequence was analysed with results shown in Figure 1. The reference frame was obtained by averaging over 200 frames. The intensity difference frame indicates the areas of candidate motion for subsequent analysis. Motion vectors were calculated as above for each pixel in the car region and plotted in Figure 2 before and after region growing. The moving region is magnified so that the individual motion vectors can be seen. A region growing map shows the separate regions P_i with the colour indicating the order of their generation according to the colour bar. $V = 10, \Phi = 30$. Approximately 90% of pixels in all the homogeneous regions (1 to 30) are assigned motion vectors.



(a) (b) (c) (d)

Figure 1: (a) Current frame; (b) next frame; (c) reference frame; (d) intensity difference frame.

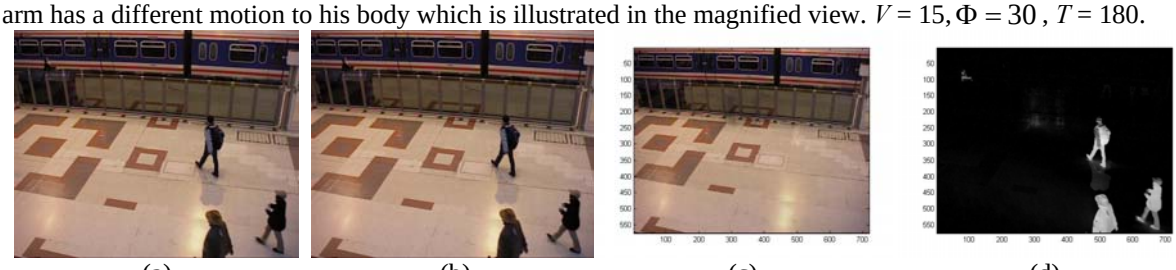


(a) (b) (c)

Figure 2: Motion vector maps (a) before and (b) after region growing; (c) region growing map.

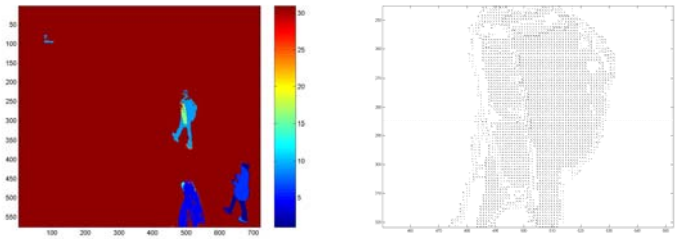
3.2 London Train Station

The results from a pair of 720x576 frames from a London Train Station sequence are shown in Figure 3. The reference frame was obtained by averaging over 1000 images taken from the video. Motion vectors are plotted in Figure 4 for the top pedestrian after region growing. A region map is shown. It is worth noting that pedestrian's arm has a different motion to his body which is illustrated in the magnified view. $V = 15, \Phi = 30$, T = 180.



(a) (b) (c) (d)

Figure 3: (a) Current frame; (b) next frame; (c) reference frame; (d) intensity difference frame.



(a) (b)

Figure 4: (a) Region growing map; (b) motion vector map for the top pedestrian.

4. Conclusions and future work

An attention based region growing algorithm has been proposed for motion detection, estimation and segmentation. The method was illustrated on various video data with a stationary background both indoors and outdoors. The proposed algorithm not only obtains motion estimation, but it also achieves a high motion vector assigning rate for motion segmentation. Furthermore the different motion regions extracted on individual objects can be used to investigate more detailed object behaviour in the scene. The method does not require a training stage or prior knowledge of the objects to be tracked.

Future work will be carried out on wider range of data with particular emphasis on tracking parts of non-rigid objects possessing different motions.

Acknowledgments

The project is sponsored by European Commission Framework Programme 6 Network of Excellence MUSCLE (Multimedia Understanding through Semantics, Computation and Learning) [16].

References

[1] H.T. Nguyen, M. Worring, and A. Dev, “Detection of moving objects in video using a robust motion similarity measure”, IEEE Trans. IP, vol. 9, pp. 137-141, 2000.
[2] M. Gelgon and P. Bouthemy, “Determining a structured spatio-temporal representation of video content for efficient visualisation and indexing”, ECCV, pp. 595-609, 1998
[3] M. Irani and P. Anandan, “Video indexing based on mosaic representations”, Proc. IEEE, vol. 86, pp. 905-921, 1998.
[4] J. Wang and Z.N. Li, “Kernel-based multiple cue algorithm for object segmentation”, SPIE, pp. 462-473, 2000.
[5] R. Piroddi and T. Vlachos, “Multiple-feature spatiotemporal segmentation of moving sequences using a rule-based approach”, BMVC, pp. 353-362, 2002.
[6] J.Y.A. Wang and E.H. Adelson, “Representing moving images with layers”, IEEE Trans. IP, vol. 3, pp. 625-638, 1994.
[7] M. Nicolescu and G. Medioni, “Motion segmentation with accurate boundaries-a tensor voting approach”, CVPR, vol. 1, pp. 382-389, 2003
[8] R. Montoliu and F. Pla, “An iterative region growing algorithm for motion segmentation and estimation”, International Journal of Intelligent Systems, vol. 20, pp. 577-590, 2005.
[9] N. Vasconcelos and A. Lippman, “Empirical Bayesian motion segmentation”, IEEE Trans. PAMI, vol. 23, pp. 217-221, 2001.
[10] A-R. Mansouri and J. Konrad, “Motion segmentation with level sets”, ICIP, vol. 2, pp. 126-130, 1999
[11] M. Chang, A. Tekalp, and M. Sezan, “Simultaneous motion estimation and segmentation”, IEEE Trans. IP, vol. 6, no.9, pp. 1326-1333. 1997
[12] S. Zhang and F.W.M. Stentiford, “An attention based method for motion detection and estimation”, Workshop on Computational Attention and Applications, ICVS, Bielefeld, Germany, 2007.
[13] S. Zhang and F. W. M. Stentiford, “Motion detection using a model of visual attention”, ICIP, San Antonio, USA, 2007.
[14] F.W.M. Stentiford, “An estimator for visual attention through competitive novelty with application to image compression”, Picture Coding Symposium, pp. 101-104, 2001.
[15] Performance Evaluation of Tracking and Surveillance (PETS), http://ftp.pets.rdg.ac.uk.
[16] Multimedia Understanding through Semantics, Computation and Learning, 2005. EC 6th Framework Programme, FP6-507752, http://www.muscle-noe.org/.