MOTION SEGMENTATION USING REGION GROWING AND AN ATTENTION BASED ALGORITHM

Shijie Zhang, Fred Stentiford

Department of Electronic and Electrical Engineering
University College London, Adastral Park Campus, Ross Building
Martlesham Heath, Ipswich, IP5 3RE, UK
{j.zhang, f.stentiford}@adastral.ucl.ac.uk
+44 (0) 1473 635199

Keywords: motion segmentation, motion estimation, object tracking, visual attention, region growing

Abstract

A novel two-stage framework for motion segmentation under stationary background conditions is proposed in this paper. The first stage uses an attention based method to extract motion information. The second stage then applies a region growing technique to motion 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 identifies sets of contiguous pixels that are moving continuously through the sequence of video frames. It is a basic task for many computer vision applications, such as content-based video retrieval in digital libraries [10], object tracking [14], and object-based analysis and video coding [11,15].

Motion segmentation can be performed by either first estimating a field of motion parameters then segmenting it, or by applying joint motion estimation and segmentation. Existing motion estimation techniques include gradient based methods [2,6,8], region based matching methods [1,3], energy based methods [5] and phase based methods [4,7]. Gradient based techniques generally perform poorly in the presence of noise. Region based methods and phase based technique are computationally expensive. In addition energy based methods often involve complex implementations.

Wang and Adelson [15] proposed one of the earliest bottom-up affine-clustering motion segmentation system for the layered representation of video. It deals with object occlusions and is suited for video coding. However its performance depends largely on parameter estimation and the choice of the origin in the k-means clustering process. Vasconcelos and Lippman [13] exploited the expectation-maximization approach and introduced an empirical Bayesian procedure for simultaneous segmentation of motion fields as well as estimation of Markov Random Field prior parameters. The image segmentation is based on linear parametric motion models embedded in a probabilistic framework. The approach eliminates the need for trial-and-error strategies for parameter setting and obtains good segmentation in fewer iterations. However, the method is computational expensive, and the clustering parameter can introduce noise into the segmentation as well as affecting accuracy near region boundaries. Nguyen et al. [10] proposed a segmentation method based on a motion similarity measure and a novel two-stage merging algorithm. Motion based region similarity is achieved by a statistical test. Using this measure, the hierarchical region merging is performed and provides a good starting point for a K-means like merging algorithm. Good initial cluster centres are required. The number of objects has to be pre-specified and an initial segmentation process is also required. Wang and Li [14] presented a novel Kernel-based multiple cue (KMC) algorithm for object segmentation. First it detects and calibrates the camera motion and finds the kernel of moving objects. The segmentation then starts from the kernels which are textured regions possessing credible motion vectors. It combines multiple cues including motion, colour and texture to achieve better segmentation. The proposed algorithm works on rigid moving objects under different camera motions. It fails when the kernel detection cannot produce good results. The proposed algorithm uses MPEG motion vectors as the major cue for object segmentation, therefore relies on the accuracy of motion vectors for good performance. Multiple features were again used in [11] where a two-stage architecture for object segmentation was proposed for moving sequences. The first stage locates objects using a hierarchy of single-feature segmentation processes using motion, texture and intensity information. The second stage refines the boundaries of located objects using a combination of features based on a set of region merging rules. The proposed method locates accurate object boundaries. The technique is computational efficient because it avoids homogeneity criteria and weighting functions that competing methods require. Most of the processing stages can be implemented in parallel. The segmentation is based on mathematical morphology which requires the object to be highly textured. In [9] an iterative region growing algorithm for motion segmentation and estimation was presented. The proposed algorithm combines

temporal information using a generalized least-squares (GLS) motion estimation process, and spatial information using an iterative region growing algorithm. The model works without a priori information on number of moving regions in the scene. A static background is not required and it works on grey-scale images so that the brightness constancy assumption can be used for GLS estimator. Lack of texture and large motions pose problems. Since it uses regions of pixels objects undergoing non-rigid motion cannot be classified.

Motion segmentation requires accurate estimation of movement. We adopt an attention based approach [16] to extract motion information which is then used in a region growing and matching process. The motion segmentation framework is presented in Section 2. Results are shown in Section 3 along with some discussion. Finally, Section 4 outlines conclusions and future work.

2 Motion segmentation framework outline

The proposed framework contains two stages. The first stage uses an attention based method [16] 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 based on visual attention

Regions of static saliency have been identified using an attention method described in [12]. 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 [16] 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

$$I_x = \{|r_2 - r_1| + |g_2 - g_1| + |b_2 - b_1|\}, \quad (1)$$

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 $\boldsymbol{F(x) = a}$ 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

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

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

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

Forks are only generated which contain pixels that mismatch each other. This means that they are selected in image regions possessing high or certainly non-zero attention scores, such as on edges or other salient features as reported earlier [12].
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.

$$|F_k(\boldsymbol{x_i'}) - F_k(\boldsymbol{x_j'})| > \delta_k, \quad \textit{for some } i, j, k. \quad (4)$$

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

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

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 threshold δ_k,

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

N attempts are made to find matches and the corresponding displacements are recorded as follows:
For the *j*th 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

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

and the cumulative displacements Δ and match counts Γ as

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

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

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

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 a copy in O_{MV}'.

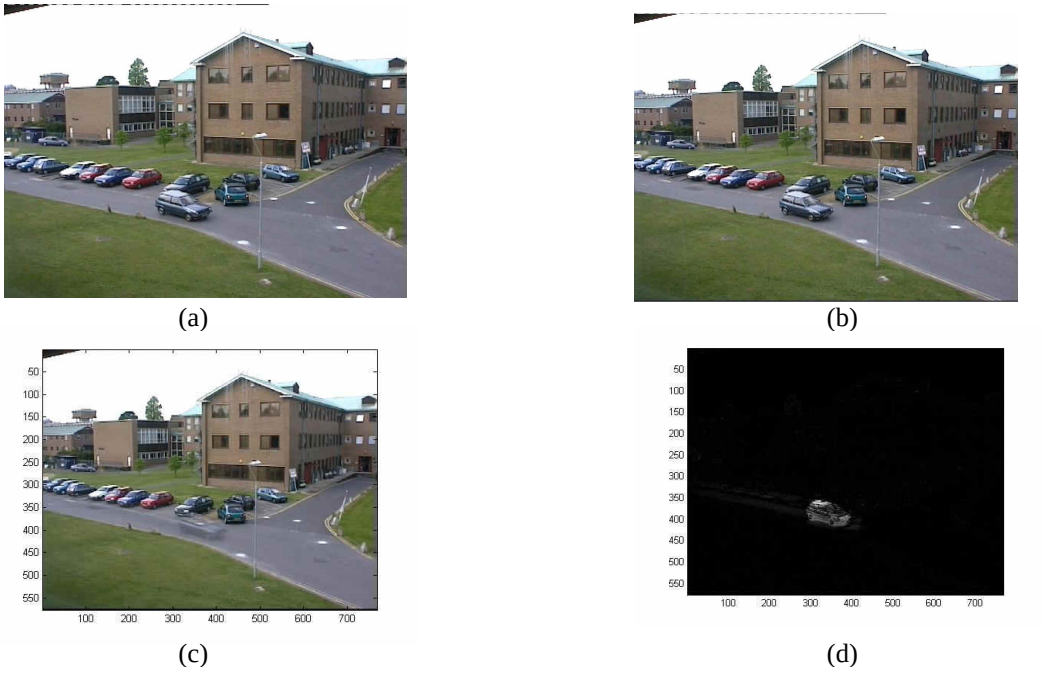


Figure 1: (a) Current frame; (b) next frame; (c) reference frame; (d) intensity difference frame.

2.2 Region growing and matching

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 [17] including road scenes from a pedestrian sequence from PETS2000, Reading University data from PETS2001 and London Train Station sequence from

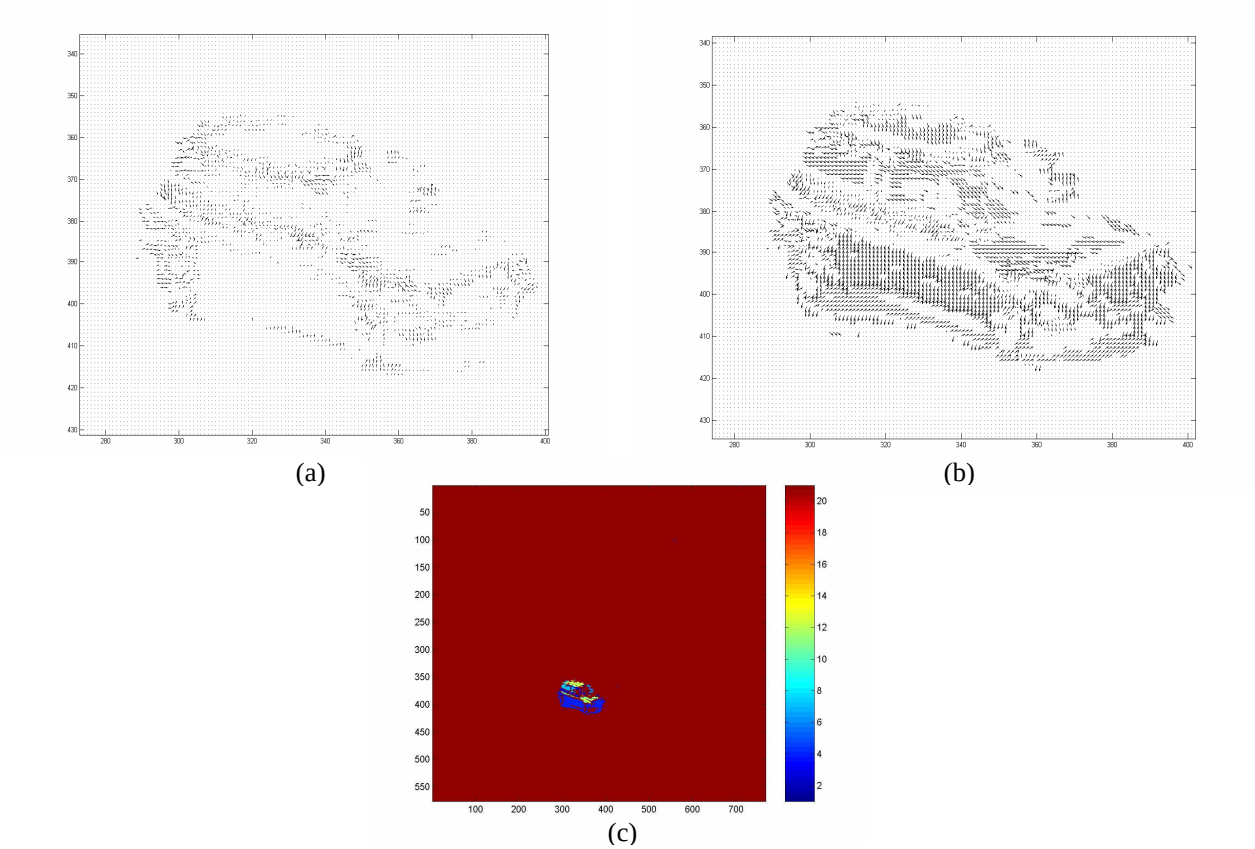


Figure 2: Motion vector maps (a) before and (b) after region growing; (c) region growing map.

PETS2006. The parameters of all experiments were M = 100, N = 10000, $\varepsilon = 1$, m = 2, $\boldsymbol{\delta}$ = (40,40,40), T = 90. The varying parameters are the view radius, V and Φ. 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, $\Phi = 20$. It is noted that the shadow area underneath the car was also assigned motion vectors. This blurs the motion boundary of the moving object and could be minimised using shadow removal techniques.

Table 1 below indicates the effectiveness of region growing in this example. Approximately 90% of pixels in all the homogeneous regions (1 to 20) are assigned motion vectors.

Region P_i	Total Pixels in region	Pixels assigned motion vectors	Percentage of pixels assigned
1	7	7	100
2	372	342	91.9
3	135	128	94.8
4	1503	1459	97
5	4	4	100
6	82	79	96.3
7	233	222	95.3
8	6	4	66.7
9	3	3	100
10	61	60	98.4
11	218	214	98.2
12	254	171	67.3
13	4	4	100
14	24	24	100
15	2	2	100
16	11	9	81.8
17	15	13	86.7
18	12	12	100
19	13	12	92.3
20	3	3	100

Table 1: Region growing statistics.

3.2 Pedestrian sequence

A pair of 768x576 frames from a pedestrian sequence was analysed with results shown in Figure 3. The reference frame was obtained by averaging over 1452 frames. The faded image of a parked car is present in the reference frame because it appeared for only part of the time during averaging. This candidate region is rejected as it is not present in the intensity difference between the current frame and the next. Motion vectors were calculated as above for each pedestrian and plotted in Figures 4 and 5. $V = 10, \Phi = 15$. Again shadow causes distortion in Fig. 5(a).

3.3 London Train Station

The results from a pair of 720x576 frames from a London Train Station sequence are shown in Figure 6. The reference frame was obtained by averaging over 1000 images taken from the video. Motion vectors are plotted in Figure 7 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. $V = 15, \Phi = 30$, $T = 180$.

4 Conclusions and future work

An attention based region growing algorithm has been proposed for motion detection, estimation and segmentation. The framework includes: an attention based approach that extracts object displacement between frames by comparing salient regions; a region growing technique that classifies motion regions according to motion information extracted; and finally a simple matching process that assigns motion vectors to the classified regions.

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 and shadow identification. More precise evaluation will also be carried out using ground truth data.

Acknowledgements

The project is sponsored by European Commission Framework Programme 6 Network of Excellence MUSCLE (Multimedia Understanding through Semantics, Computation and Learning) [18].

References

[1] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion", IJCV, vol. 2, pp. 283-310, (1989).

[2] J.R. Bergen, P.J. Burt, R. Hingorani, S. Peleg, "A three-frame algorithm for estimating two-component image motion", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 14, pp. 886-896, (1992).

[3] M.J. Black, P. Anandan, "Robust dynamic motion estimation over time", CVPR, pp. 296-302, (1991).

[4] D.J. Fleet, A.D. Jepson, "Computation of component image velocity from local phase information", IJCV, vol. 5, pp. 77-104, (1990).

[5] D.J. Heeger, "Optical flow using spatiotemporal filters", IJCV, vol. 1, pp. 279-302, (1988).

[6] B.K.P. Horn, B.G. Schunck, "Determining optical flow", Artificial Intelligence 17, pp. 185-204, (1981).

[7] A.D. Jepson, D.J. Fleet, T.F. El-Maraghi, "Robust online appearance models for visual tracking", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, pp. 1296-1311, (2003).

[8] B.D. Lucas, T. Kanade, "An iterative image registration technique with an application to stereo vision", Proc. Of Imaging Understanding Workshop, pp. 121-130, (1981).

[9] R. Montoliu, F. Pla, "An iterative region growing algorithm for motion segmentation and estimation", International Journal of Intelligent Systems, vol. 20, pp. 577-590, (2005).

[10] H.T. Nguyen, M. Worring, A. Dev, "Detection of moving objects in video using a robust motion similarity measure", IEEE Trans. Image Processing, vol. 9, pp. 137-141, (2000).

[11] R. Piroddi, T. Vlachos, "Multiple-feature spatiotemporal segmentation of moving sequences using a rule-based approach", BMVC, pp. 353-362, (2002).

[12] F.W.M. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression", Picture Coding Symposium, pp. 101-104, (2001).

[13] N. Vasconcelos, A. Lippman, "Empirical Bayesian motion segmentation", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, pp. 217-221, (2001).

[14] J. Wang , Z.N. Li, "Kernel-based multiple cue algorithm for object segmentation", SPIE, pp. 462-473, (2000).

[15] J.Y.A. Wang, E.H. Adelson, "Representing moving images with layers", IEEE Trans. Image Processing, vol. 3, pp. 625-638, (1994).

[16] S. Zhang, F.W.M. Stentiford, "An attention based method for motion detection and estimation", Workshop on Computational Attention and Applications, ICVS, (2007).

[17] Performance Evaluation of Tracking and Surveillance (PETS), http://ftp.pets.rdg.ac.uk.

[18] Multimedia Understanding through Semantics, Computation and Learning, 2005. EC 6th Framework Programme, FP6-507752, http://www.muscle-noe.org/.

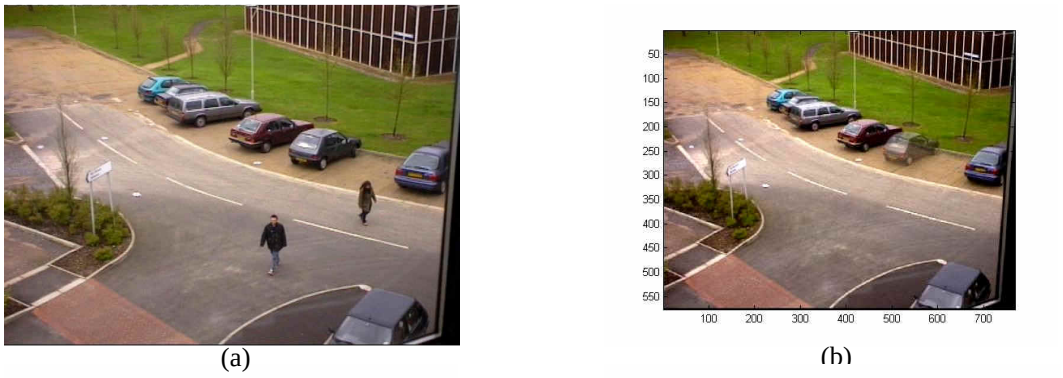


Figure 3: (a) Current frame; (b) reference frame; (c) thresholded intensity difference between current and reference frame; (d) intensity difference between current and next frames

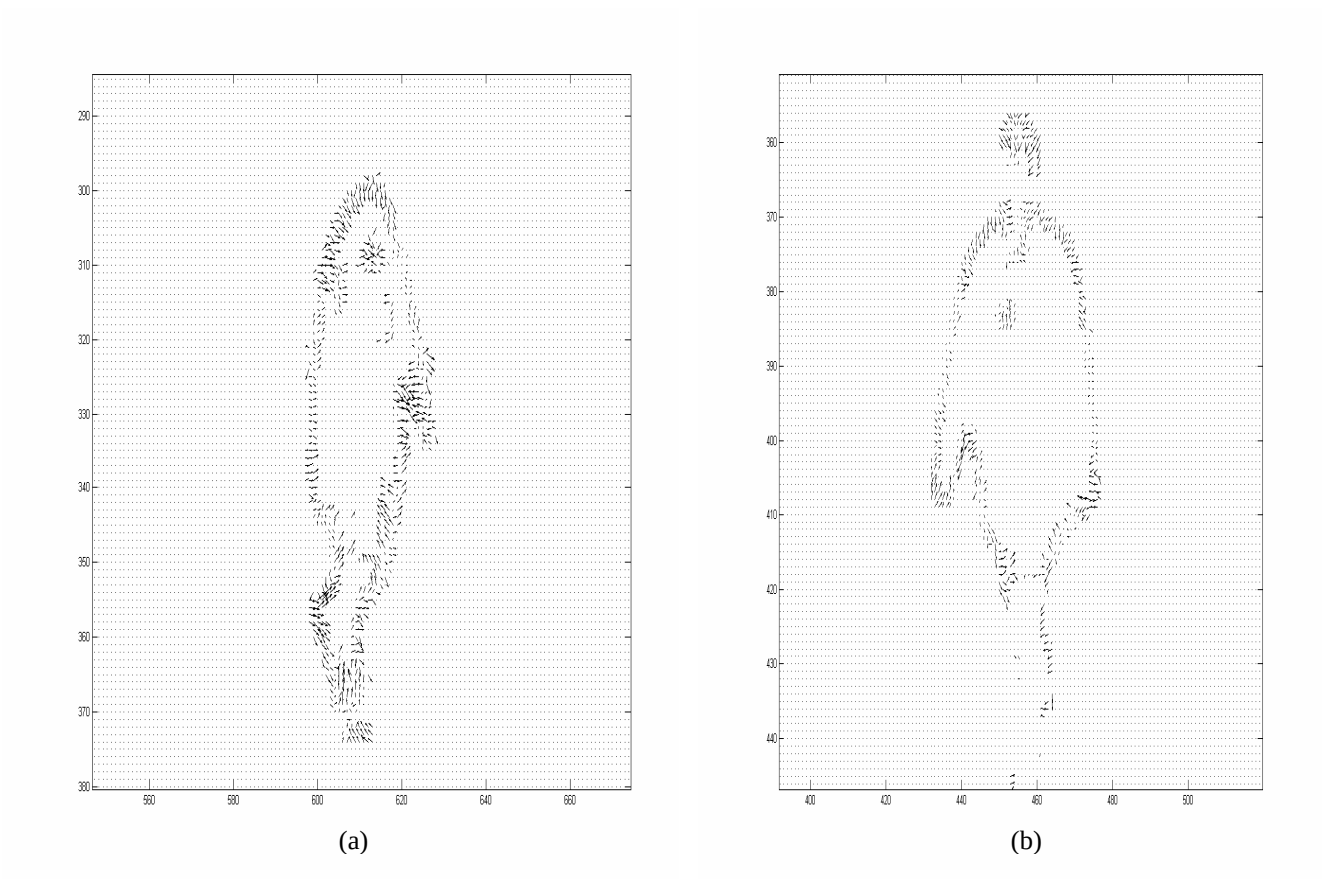
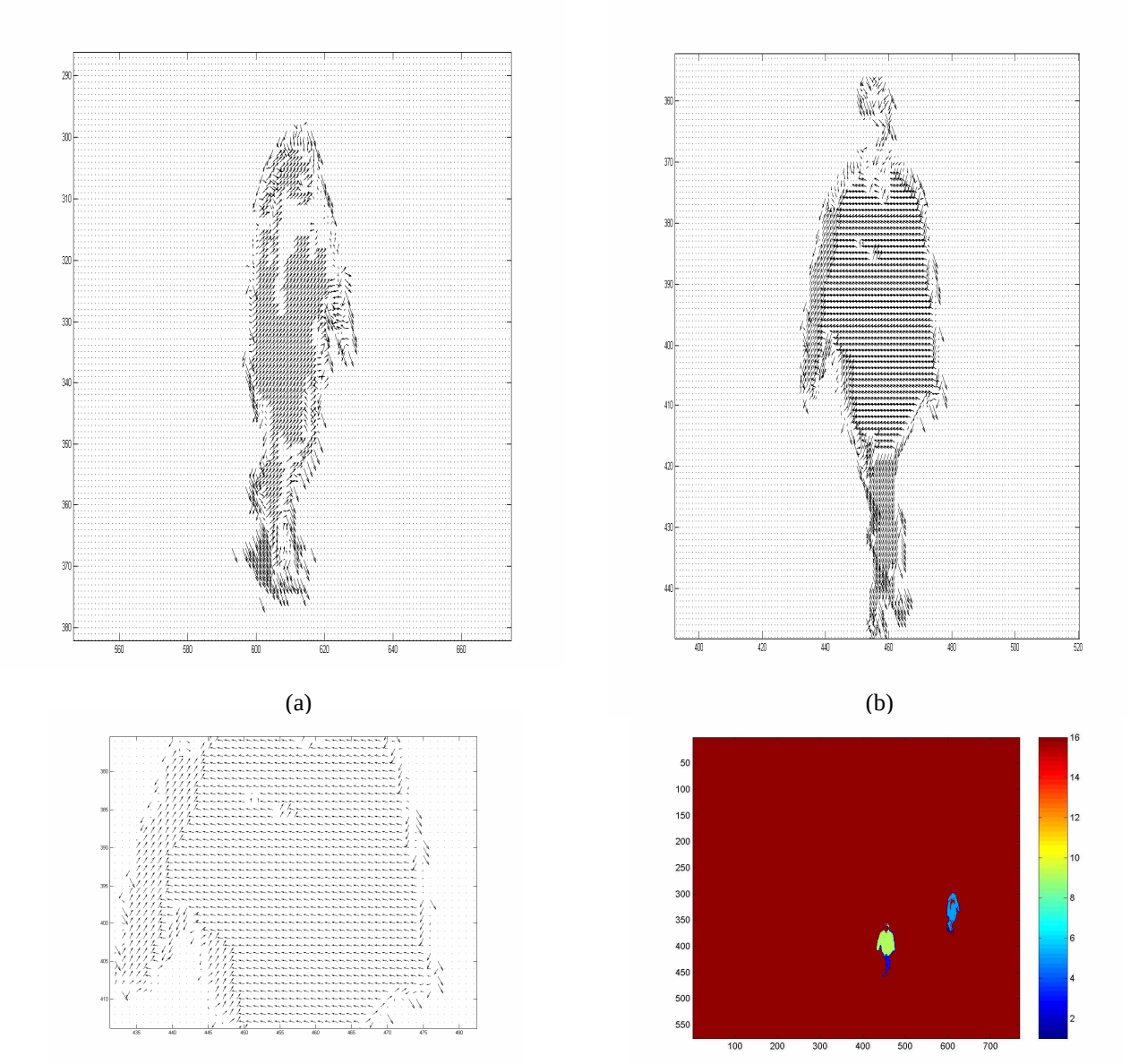


Figure 4: Motion vector maps before region growing. (a) pedestrian 1; (b) pedestrian 2.



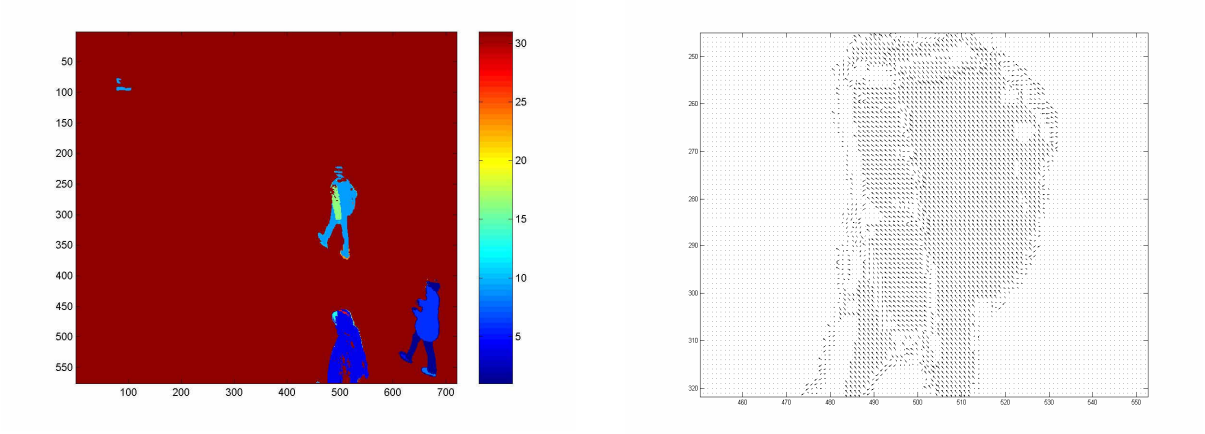
(a) (b)

(c) (d)

Figure 5: Motion vector maps after region growing. (a) pedestrian 1; (b) pedestrian 2; (c) magnified view of (b) on the pedestrian's body; (d) region growing map.

(a) (b)

Figure 6: (a) Current frame; (b) intensity difference between current and reference frames.



(a) (b)

Figure 7: (a) Region growing map; (b) motion vector map for the top pedestrian.