This is a preliminary version of an article published in
Proc. of IEEE International Conference on Image Processing (ICIP), pp. 73 – 76, Thessaloniki, Greece, October 2001
by Gerald Kühne, Joachim Weickert, Oliver Schuster, Stephan Kopf

A TENSOR-DRIVEN ACTIVE CONTOUR MODEL FOR MOVING OBJECT SEGMENTATION

Gerald Kühne[*]*, Joachim Weickert, Oliver Schuster, Stephan Richter*

University of Mannheim
Department of Mathematics and Computer Science
68131 Mannheim, Germany

ABSTRACT

In this paper we propose an approach to the segmentation of video objects based on motion cues. Motion analysis is performed by estimating local orientations in the spatio-temporal domain using the three-dimensional structure tensor.

These estimates are integrated as an external force into an active contour model, thus stopping the evolving curve when it reaches the moving object's boundary. To enable simultaneous detection of several objects, we reformulate the tensor-based active contour model using the level-set technique. In addition, a contour refinement technique has been developed to better approximate the real boundary of the moving object.

We provide promising experimental results calculated on real-world video sequences widely used within the computer vision community.

1. INTRODUCTION

Algorithms for automatic segmentation of objects from a video sequence are required for a variety of applications ranging from video compression to object recognition. For instance, the MPEG-4 video coding standard [1] provides functionality for content-based access. Video information can be encoded in a number of arbitrarily shaped video object planes representing instances of video objects. In addition, algorithms for high-level vision tasks such as shape-based object recognition [2] depend on information with regard to object outlines.

Motion cues available in video sequences facilitate the segregation of moving objects from the background. Various approaches have been proposed in this field. However, many of them determine basic motion parameters on the basis of only two consecutive frames. Hence, these techniques are sensitive to noise and require appropriate compensation methods. In [3] an edge map is calculated from the difference image of two frames. Mech and Wollborn [4] and Paragios and Deriche [5] employ a statistical framework, while Meier and Ngan [6] perform a connected component analysis on the observed inter-frame differences.

In our approach, we analyze motion by estimating local orientations in the spatio-temporal domain using the three-dimensional structure tensor, thus exploiting motion information from a space-time continuum. The estimates gained from the structure tensor are integrated as an external force into a level-set based active contour model. This model allows the simultaneous detection of several objects and also enables closure of holes and gaps in the motion detection result.

The remainder of the paper is organized as follows: Section 2 introduces the tensor-based motion detection algorithm. Sections 3 and 4 describe the integration of the structure tensor into the active contour model and a contour refinement technique. Section 5 presents experimental results. Finally, Section 6 offers concluding remarks.

2. TENSOR-BASED MOTION DETECTION

Within consecutive frames stacked on top of each other, a video sequence can be represented as a three-dimensional volume with two spatial (x, y) and one temporal (z) coordinates. From this perspective, motion can be estimated by analyzing orientations of local gray-value structures [7]. Under the assumption of a non-varying illumination, gray values remain constant in the direction of motion. Thus, stationary parts of a scene result in lines of equal gray values in parallel to the time axis. Moving objects, however, cause iso-gray-value lines of different orientations. Figure 1 illustrates this observation.

Consequently, moving and static parts on the image plane can be determined from the direction of minimal gray value change in the spatio-temporal volume. This direction can be calculated as the direction $\mathbf{n}$ being as much perpendicular to all gray-value gradients in a 3D local neighborhood Ω. Thus, we minimize

$$\int_{\Omega} (\nabla_3 I(x, y, z)\mathbf{n})^2 dx\ dy\ dz \qquad (1)$$

[*]Send correspondence to: kuehne@informatik.uni-mannheim.de

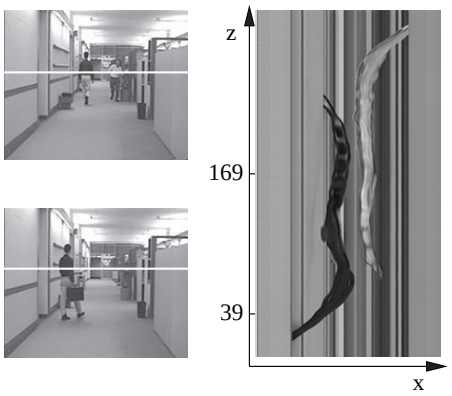

Fig. 1. Local orientation of image structures. Left: Frame 169 (top) and frame 39 (bottom) of the "hall and monitor" sequence. Right: Slice of the corresponding spatio-temporal volume taken at the horizontal line marked by the white lines in the single frames.

where $I(x, y, z)$ denotes the three-dimensional volume and $\nabla_3 := (\partial_x, \partial_y, \partial_z)$ the spatio-temporal gradient.

As described in [7, 8] minimizing Equation 1 is equivalent to determining the eigenvector of the minimum eigenvalue of the 3D structure tensor

$$\mathbf{J} = \begin{bmatrix} J_{xx} & J_{xy} & J_{xz} \\ J_{xy} & J_{yy} & J_{yz} \\ J_{xz} & J_{yz} & J_{zz} \end{bmatrix} \tag{2}$$

where $J_{pq}, p, q \in \{x, y, z\}$ are calculated within a local neighborhood Ω from

$$J_{pq}(x, y, z) = \int_{\Omega} \partial_p I(x', y', z') \partial_q I(x', y', z') dx'\, dy'\, dz'. \tag{3}$$

By analyzing the three eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ of the structure tensor, we can classify the local neighborhood's motion. All three eigenvalues equal to zero indicates an area of constant gray values, therefore no motion can be detected. If $\lambda_1 > 0$ and $\lambda_2 = \lambda_3 = 0$, gray values change only in one direction. Consequently, due to the correspondence problem we can calculate only normal velocity. Real motion can be calculated if gray values remain constant in only one direction, hence, $\lambda_1 > 0, \lambda_2 > 0$ and $\lambda_3 = 0$. Finally, if all three eigenvalues are greater than zero, we cannot determine the optical flow.

In real-world video sequences, however, it is impractical to compare the eigenvalues to zero since due to noise in the sequence small gray-value changes always occur. Thus, we introduce normalized coherence measures c_t and c_s that quantify the certainty of the calculations. Note that our measures deviate from those defined in [8] and avoid the discontinuity problem [9]. The coherence measure c_t indicates

Fig. 2. Tensor-based motion detection on frame 12 of the Hamburg taxi sequence. Motion pixels are marked white. From left to right: (a) original frame, (b) motion detection with $|\Omega| = 3^3$, (c) motion detection with $|\Omega| = 7^3$. $C = 5$ was used in all calculations.

whether a reliable motion calculation is possible and is defined by

$$c_t = \begin{cases} 0 & \lambda_1 = \lambda_3, \\ \exp\left(\frac{-C}{|\lambda_1 - \lambda_3|}\right) & \text{else} \end{cases} \tag{4}$$

where $C > 0$ denotes a contrast parameter. Areas with $(|\lambda_1 - \lambda_3|) \ll C$ are regarded as almost constant local neighborhoods [9]. A value of c_t near 1.0 indicates that $\lambda_1 \gg \lambda_3$, therefore a reliable motion calculation can be performed. The opposite is true if the c_t value approaches zero.

The coherence measure c_s,

$$c_s = \begin{cases} 0 & \lambda_2 = \lambda_3, \\ \exp\left(\frac{-C}{|\lambda_2 - \lambda_3|}\right) & \text{else} \end{cases}, \tag{5}$$

indicates whether normal or real motion can be determined. Values near one allow the calculation of real motion. Otherwise, only normal velocities can be specified.

Hence, our motion detection scheme works as follows. At each position in the video sequence the structure tensor and the coherence measures are calculated. If c_t is near 1.0, the motion vector is determined in accordance with c_s. Then, under the assumption of a static camera, the position is marked as a "motion pixel" if the 2D velocity v, i. e., the norm of the motion vector, exceeds a certain threshold T_v, e.g. $T_v > 0.1$ pixel/frame.

In the event of a moving camera, a global camera motion estimation has to be performed first. It is then possible to compare the motion vector determined from the structure tensor to the vector resulting from the the global camera parameters [6].

Figure 2 illustrates the perfomance of our motion detection scheme on frame 12 of the Hamburg taxi sequence. The sequence contains four moving objects: the taxi in the middle, a car on the left, a van on the right, and a pedestrian in the upper left corner of the image. Note that an increase of the neighborhood's size significantly reduces the amount of noise in the image without changing the parameter C.

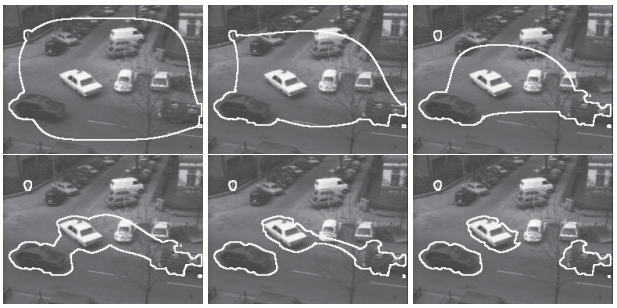

Fig. 3. Tensor-driven geodesic active contour. From top left to bottom right: contour after 3000, 6000, 9000, 12000, 15000, 17392 iterations. Constant force $c = 0.02$, $\epsilon = 0.25$.

The tensor approach is able to detect motion areas reliably, though some parts of the van are left out due to low contrast. However, we observe two shortcomings that are inherent to this approach. First, due to areas of constant gray values within the moving objects, we do not receive dense motion vector fields . Second, the tensor fails to provide the true object boundaries accurately since the calculations within the neighborhood Ω blurs motion information across spatial edges.

3. TENSOR-DRIVEN ACTIVE CONTOUR MODEL

Tensor-based motion detection identifies regions of motion. However, these regions need neither be connected nor form semantically meaningful objects. Consequently, we need a grouping step that integrates neighboring regions into objects while closing gaps and holes.

Widely used within this context are active contour models. Basically, a planar parametric curve $\mathcal{C}(s)$ placed around image parts of interest evolves under smoothness control (internal energy) and the influence of an image force (external energy).

In the classical explicit snake model [10] the following functional is minimized

$$\oint_{\mathcal{C}(s)} \left(\alpha|\mathcal{C}'(s)|^2 + \beta|\mathcal{C}''(s)|^2 - \gamma|\nabla I(\mathcal{C}(s))|^2\right) ds \quad (6)$$

where the first two terms control the smoothness of the planar curve, while the third attracts the contour to high gradients of the image.

To obtain topological flexibility that allows the simultaneous detection of multiple objects, we employ geodesic active contours [11, 12]. The basic idea is to embed the initial curve as a zero level set into a function $u : \mathbb{R}^2 \rightarrow \mathbb{R}$, i. e., $\mathcal{C}$ is represented by the set of points $\mathbf{x}_i$ with $u(\mathbf{x}_i) = 0$, and to evolve this function under a partial differential equation.

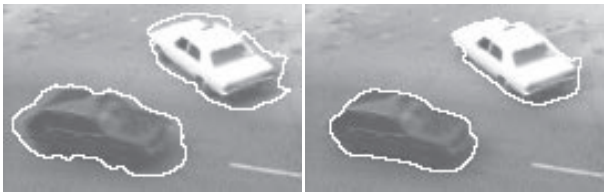

Fig. 4. Contour refinement. Left: motion-based segmentation, right: motion-based segmentation with contour refinement (445 iterations, $\tilde{C} = 1.5$).

Using a modified energy term this results in the image evolution equation [11, 12]

$$\frac{\partial u}{\partial t} = g(I)(c + \kappa)|\nabla u| + \nabla u \cdot \nabla g \quad (7)$$

where κ denotes the curvature of a level set, $\nabla := (\partial x, \partial y)$ is the spatial gradient, c adds a constant force for faster convergence, and g represents the external image-dependent force or stopping function.

By defining an appropriate stopping function g we can integrate the tensor-based motion detection into the model. Choosing $g(I) = \hat{s}(I)$ where $\hat{s}$ is a smoothed version of

$$s(I(x,y,t)) = \begin{cases} 1 & c_t(x,y,t) < 1-\epsilon, \\ 1 & c_t(x,y,t) \geq 1-\epsilon \wedge v(x,y,t) < T_v, \\ 0 & c_t(x,y,t) \geq 1-\epsilon \wedge v(x,y,t) \geq T_v \end{cases} \quad (8)$$

stops the curve evolution ($g = 0$) upon reaching positions that coincide with "motion pixels". Remember that v denotes the 2D velocity and T_v the velocity threshold.

Figure 3 depicts the evolution of the tensor-driven geodesic active contour. The contour succeeds in splitting up and detecting the four different moving objects.

4. CONTOUR REFINEMENT

In order to improve the segmentation results, we employ a refinement procedure based not on motion information but on the gradient values within a single frame. As can be seen in Figure 4 (left), the motion-based segmentation detects regions that are slightly larger than the moving objects.

Thus, we restart the image evolution process using the result from the motion-based segmentation as the zero level set. However, this time a stopping function $\tilde{g}$ based on the spatial gradient is used:

$$\tilde{g}(I) = \frac{1}{1 + |\nabla \hat{I}|^2/\tilde{C}^2} \quad (9)$$

Here, $\tilde{C}$ is a contrast parameter that diminishes the influence of small gradient values. Figure 4 depicts the performance of the refinement procedure.

Fig. 5. Top: Examples from the "hall and monitor" sequence, bottom: examples from the "Hamburg taxi" sequence.

5. EXPERIMENTAL RESULTS

We applied the proposed motion detection scheme to the real-world sequences "Hamburg taxi" and "hall and monitor". While the first includes mainly rigid objects, the latter contains non-rigid objects. In both cases, the structure tensor calculations were carried out using the parameters $|\Omega| = 7^3, C = 5, \epsilon = 0.25$, and $T_v = 0.1$.

Figure 5 depicts examples from the selected sequences. The moving objects are detected reliably and the active contour approximates the real object boundaries during the refinement step quite well.

6. CONCLUSION AND OUTLOOK

We presented an approach to segmenting objects based on motion cues. In particular, we employed the 3D structure tensor to derive information from the spatio-temporal volume constructed by the sequence of frames. This improves the motion detection results significantly compared to calculations based on only two consecutive frames.

The integration of the 3D structure tensor as the external force within the geodesic active contour model results in a reliable scheme for the simultaneous detection of multiple moving objects. The subsequently employed contour refinement enables the accurate segmentation of the objects in question.

There are, however, several areas that require further development. First, the contour refinement is based on rather simple gradient calculations. The use of an enhanced edge detection scheme should improve the results. Second, and more relevant, the problem of large velocities should be addressed within a multi-resolution framework.

7. REFERENCES

[1] ISO/IEC 14496-2, "Information technology – coding of audio-visual objects – part 2: Visual", 1999.

[2] S. Abbasi, F. Mokhtarian, and J. Kittler, "Enhancing CSS-based shape retrieval for objects with shallow concavities", *Image and Vision Computing*, vol. 18, no. 3, pp. 199–211, 2000.

[3] C. Kim and J. Hwang, "A fast and robust moving object segmentation in video sequences", in *Int. Conf. Image Processing (Kobe, Japan)*, 1999, pp. 131–134.

[4] R. Mech and M. Wollborn, "A noise robust method for segmentation of moving objects in video sequences", in *Int. Conf. Acoustics, Speech and Signal Processing (Munich, Germany)*, 1997, pp. 2657–2660.

[5] N. Paragios and R. Deriche, "Geodesic active contours and level sets for the detection and tracking of moving objects", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 3, pp. 266–280, March 2000.

[6] T. Meier and K. N. Ngan, "Extraction of moving objects for content-based video coding", in *Proc. SPIE, VCIP*, 1999, vol. 3653, pp. 1178–1189.

[7] J. Bigün, G. H. Granlund, and J. Wiklund, "Multidimensional orientation estimation with applications to texture analysis and optical flow", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, pp. 775–790, 1991.

[8] H. Haußecker and B. Jähne, "A tensor approach for precise computation of dense displacement vector fields", in *Proc. Mustererkennung*, E. Paulus and F.M. Wahl, Eds., Berlin, 1997, Informatik Aktuell, pp. 199–208, Springer.

[9] J. Weickert, "Coherence-enhancing diffusion of colour images", *Image and Vision Computing*, vol. 17, pp. 201–212, 1999.

[10] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models", *Int. J. Comp. Vision*, vol. 1, pp. 321–331, 1988.

[11] S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi, "Conformal curvature flows: from phase transitions to active vision", *Arch. Rat. Mech. Anal.*, vol. 134, pp. 275–301, 1996.

[12] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours", *Int. J. Comp. Vision*, vol. 22, no. 1, pp. 61–79, 1997.