52. Internationales Wissenschaftliches Kolloquium
Technische Universität Ilmenau
10.-13. September 2007

in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische Universität Ilmenau, Germany, Vol. II, pp. 327-332, Universitätsverlag 2007

S. Hellbach / B. Lau / J. P. Eggert / E. Körner / H.-M. Gross

Multi-Cue Motion Segmentation

1 Introduction

This paper presents an approach to multi-object image segmentation based on object motion using Markov Random Fields[1]. To support the information gained from motion and to achieve robustness, several additional visual cues extracted from the image data are integrated. Depth information gained from stereo disparity is included to maintain segmentation in case an segmented object stops moving. Motion is estimated with a correspondence matching scheme. The approach differs from regular optical flow in the way that rich matching results are used for segmentation rather than only the best matches. The representation of segmented regions is realized implicitly as labeling on a 2D lattice.

Motion segmentation is a key to many modern image processing applications. In video compression algorithms, the analysis of motion and regions with coherent motion helps to drastically reduce the amount of information that has to be stored and transmitted for each frame [11]. Motion segmentation and motion understanding, for example, plays an essential role in detecting and/or avoiding obstacles in vehicles or with a mobile robot.

The rest of this paper is organized as follows: while Sect. 2 provides a short overview of the work done in the field of image segmentation and relations to our approach, Sect. 3 describes the architecture of the proposed system. Its evaluation is presented in Sect. 4. Finally, the paper concludes with Sect. 5.

2 State of the Art

Recent research in the area of motion segmentation focuses on feature-based motion estimation, approaches using level set methods [4], and on multi-cue segmentation on Markov Random Fields [6]. Optimization on Markov Random Fields is an established approach to segmentation [9]. It permits to let motion estimation be part of the segmentation process in one single optimization framework.

When Markov Random Fields are used in combination with optimization methods like iterated conditional modes (ICM) [2], the initialization of the field configuration and the label parameters has a strong influence on the time needed for convergence as well as on the quality of the result. In multi-frame applications, results from the last frame can be used for initialization [3]. Our approach uses motion estimates that fulfill certain quality criteria, as well as results from the previous frame.

[1] This paper reflects the results of Boris Lau's diploma thesis, available online at `http://www.borislau.de/computerscience/publications/`

in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische
Universität Ilmenau, Germany, Vol. II, pp. 327-332, Universitätsverlag 2007

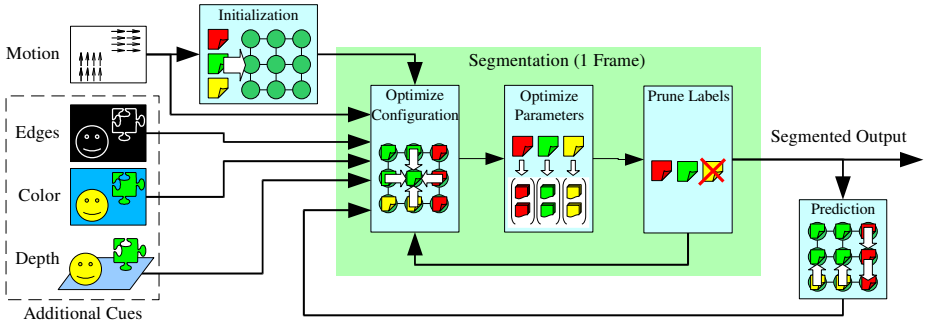


Figure 1: Schematic overview of the framework's architecture. Multiple visual cues are incorporated by the segmentation framework. The data resulting from the segmentation in the current time step is used as initialization for the following by applying the motion information for each segment.

In state-of-the-art optimization-based motion segmentation systems with an implicit representation, motion information is included either in the form of spatio-temporal image gradients [9], or as optical flow [12]. Our system transfers the approach to use sum of squared differences (SSD) surfaces instead of just the best matches [8] to motion segmentation by directly integrating SSD surfaces into a cost function.
Our system as well as other approaches (e.g. [5]) alternately estimates the labeling and the motion parameters of the labels using least squares fitting as in [3].
Correct correspondence matches cannot be found for occlusions, i.e. areas that are freshly covered or uncovered by a moving object. Stiller [10] proposed to use the displacement field of the previous image frame. The estimation of a binary (dis)occlusion field has been done using the residual *motion-compensation error* as a marker [12].
Whenever pure motion information is not sufficient for robust segmentation, the integration of supplementary visual modalities is appealing. In most cases the mutliple cues are combined in a sequential way (e.g. [1]), but this does not exploit the supplementary nature of the cues [6]. However, some work has been done that follows the concept of combining multiple cues in one cost function, for example by using edges [12] to enhance segment boundaries, by evaluating uniformity of both color and motion with adaptive weights [6].In contrast to [7], who utilizes disparity information for occlusion detection, our system uses it directly for segmentation.

3 System Architecture

The algorithm presented in this paper is designed following the concept of Fig. 1. Motion, edges, color, and depth information are the visual cues used in this paper. For the actual segmentation algorithm we have chosen a Markov Random Field (MRF) framework as fundament. As done by other authors, we are alternating two different optimization steps during the iterative optimization for one image frame.

in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische
Universität Ilmenau, Germany, Vol. II, pp. 327-332, Universitätsverlag 2007

Motion estimation: Motion is estimated by rating correspondences, using the sum of squared differences metric $SSD(x, y, \Delta x, \Delta y)$. Classic motion estimation approaches select the displacement with the minimum SSD value from the SSD surface as the motion estimate and discard the rest. In many cases this is not appropriate, for example, if no or only low contrast is present at all or only in one direction in the image patch, also referred to as "aperture problem". Instead, the whole SSD surface is used as input data for the optimization procedure during segmentation. This way, information from ambiguous matching surfaces is used in a sensible way.

To detect areas of occlusion where the image changes cannot be explained by motion, we use the minimum of an SSD surface as a measure for its validity: only if there is a low minimum in an SSD surface, the respective motion can account for the brightness changes in the correlation window at that particular location. If the minimum is higher than a certain threshold τ_{occl}, the matching is assumed to be bogus. Such SSD surfaces are completely set to zero so they do not affect motion segmentation.

Motion segmentation: The segmentation of an image is specified by a labeling (configuration) of a Markov Random Field with one site (vertex in the Markov Random field graph) $s(x, y)$ for each pixel. The value assigned to a site is its label $l \in \{1, ..., L\}$. The number of labels L can change during optimization and from frame to frame. Each label l is associated with a displacement $\Delta x_l, \Delta y_l$ and other optional features like a normalized color histogram $c_l(h, s)$ of hue and saturation values or the average depth d_l of all sites labeled with l. Thus, sites with the same label are considered to show the same translational motion, and if using the optional descriptive features, similar coloring and depth. So, the framework bootstraps the object knowledge from the presented scene.

Optimization of site configuration: How well a label l fits the local image properties is expressed with the *fidelity* term FID. Good correspondence matches have a low SSD value, and similar colors have a high value in the histogram. The weights $\alpha_{col} \geq 0$ and $\alpha_{dep} \geq 0$ control the influence of the color cue and the depth cue respectively:

$$FID(x, y, l) = SSD(x, y, \Delta x_l, \Delta y_l) - \alpha_{col} \cdot c_l(H(x, y), S(x, y)) + \alpha_{dep} \cdot (d(x, y) - d_l)^2 \ . \quad (1)$$

Furthermore, to introduce smoothness constraints, a regularization term (2) is formulated for each site s that assigns a penalty for each adjacent site s' in a 4-neighborhood that has a different label than s. We use the inverted Kronecker-Delta-function $\bar{\delta}(a, b)$ which is 0 if $a = b$ and 1 otherwise. The penalty for a neighboring site s' is reduced if a brightness edge goes through exactly one of the two sites in a pair (s, s') with image coordinates (x, y) and (x', y') respectively. With $\mathcal{N}_4$ being the 4-connected neighborhood, the regularization is defined as follows:

$$REG(x, y, l) = \sum_{(x', y') \in \mathcal{N}_4} \bar{\delta}\left(l, s\left(x', y'\right)\right) \cdot \left(1 - \alpha_{edg} \cdot \bar{\delta}\left(e\left(x, y\right), e\left(x', y'\right)\right)\right) \quad (2)$$

in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische
Universität Ilmenau, Germany, Vol. II, pp. 327-332, Universitätsverlag 2007

The optimization of the labeling of the sites is done with greedy local optimization called *Iterated Conditional Modes* (ICM), which is identical to *Simulated Annealing* with a minimal temperature from the very beginning [2]. The update in each iteration is done for all sites in a random order, always selecting the label for a site which minimizes the sum of fidelity and regularity for that site.

Optimization of labels: The motion parameters of the labels as well as the color information have to be updated along with the configuration in each iteration step. For each label l the displacement values are chosen that belong to the minimum in the sum of all SSD surfaces of sites with the same label l. Labels that are not assigned to any site are deleted from the list. Labels with identical motion parameters are unified. This way the total number of labels L can decrease during one iteration.

For color representation, the normalized 2D (10×10) hue-saturation color histogram $c_l(h, s)$ of a label l is computed by accumulating the hue and saturation values $H(x, y)$ and $S(x, y)$ from all sites with label l. The depth information d_l is represented as an average over all pixels with the same label l, i.e. pixels that belong to the same segment.

After this optimization step is done, the information describing the segmented object is represented by the label parameters. Hence, the system is able to segment unknown objects. Furthermore, objects with smoothly changing occurence can be tracked.

Initialization and motion propagation: Besides a standard (background) label with $\Delta x = \Delta y = 0$, the initial labels are determined by searching for "good" motion estimates in the SSD surfaces. These have to meet all of the following criteria: (a) Motion needs to be clearly present, which is indicated by bad matching results for $\Delta x = \Delta y = 0$. (b) Occlusion areas are excluded, as defined for motion estimation. (c) The correspondence matches have to be unambiguous, which corresponds to a peaked minimum in an SSD surface.

When segmenting a sequence of images, the label parameters are carried over to the next frame. To account for dramatic changes or new occurrences of motion, new additional labels are included. The configuration is also carried over to the next frame as a prediction for the positions and shape of the segments: the motion estimates from the labels are applied to each site, yielding a linear motion model.

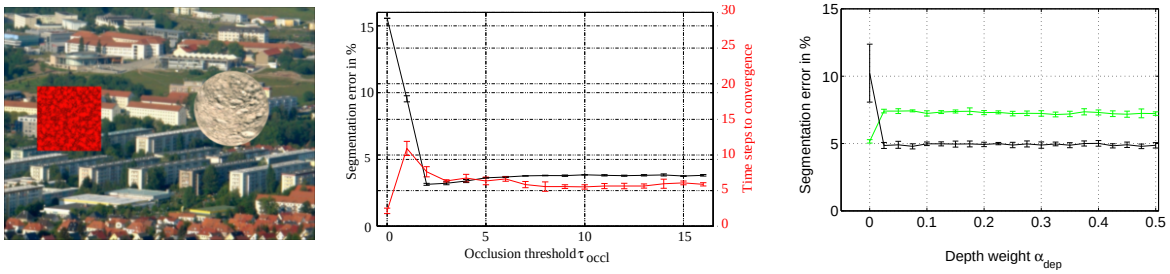
4 Experiments and Results

This section presents experiments to demonstrate the performance of our system. The evaluation is done on synthetic as well as on real world data. The synthetic data (Fig. 2a), showing two textured objects moving in front of cluttered background, is generated using POV-Ray[2]. The real world data contains standard MPEG evaluation sequences, as well as videos recorded with a VidereDesign STOC[3] stereo camera device. Due to missing

[2] `http://www.povray.org`

[3] `http://www.videredesign.com`

in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische
Universität Ilmenau, Germany, Vol. II, pp. 327-332, Universitätsverlag 2007



(a) (b) (c)

Figure 2: Experimental results using a synthetic test sequence (a) with two objects moving in front of cluttered background. The diagram (b) show the segmentation error (black) and the needed time steps (red) while changing the the occlusion threshold τ_{occl}. The plot for the depth weight α_{dep} (c) shows the results for different types of sequences (black and green) . Values are averaged over 10 trials with 10 frames each. The error bars show the standard deviation over the trials.

ground truth data the quantitative tests are performed only on the synthetic data set. Both segmentation error and number of iterations are evaluated by changing one of the system's parameters while keeping the others fixed. The segmentation error is calculated taking the percentage of pixels which were labeled the wrong segment, according to [13]. Results considering real world images are not discussed within this paper, due to a lack of space. The SSD surfaces are obtained using a 5×5 pixel correlation window, and a 15×15 pixel search window. The parameter $\alpha_{\text{reg}} = 0.5$ is determined by experimental evaluation.

In this experiment, segmentation performance is tested with different settings for the occlusion detection threshold τ_{occl}. When occlusions are not treated in particular, false segmentation occurs at frontal boundaries of moving objects. The results of this experiment are presented and discussed in Fig. 2b. If the threshold τ_{occl} is too high to be reached, the occlusion handling does not work. With adequate settings the false segments at the front border of the moving objects are suppressed. If τ_{occl} is too low, good motion estimates are discarded and the segmentation deteriorates. Further experiments use $\tau_{\text{occl}} = 4$.

To analyze the influence of the depth cue, two different scenes are regarded. The first one is our standard POV-Ray generated one. In the second one, one of the objects slows down below a detectable velocity. The results can be seen in Fig. 2c. The green plot is evaluated with the sequence containing continuous movement, while the black one shows the results with absent movement. It can be seen in the black curve, that the segmentation result becomes better with an enabled depth cue $\alpha_{\text{dep}} > 0$. Our framework is able to compensate the absent motion with the help of the depth cue. The green plot shows a slight disimprovement with the stereo cue enabled. This is due to the fact, that the depth information we gain from the stereo camera device contains fuzzy borders around each

in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische
Universität Ilmenau, Germany, Vol. II, pp. 327-332, Universitätsverlag 2007

object. So, the position of the border caused by the motion differs from the one of the depth estimation, and the system is unable to find the right one.

5 Conclusion

We have presented a motion segmentation system that integrates additional cues like edges, color, and depth information in one optimization scheme. Correspondence-based motion information is incorporated with full SSD surfaces instead of best matches only. This way, data from evenly good matches in ambiguous cases is not discarded.

Results have been presented for segmentation on rendered scenes. We have shown the advantage of multi-cue segmentation over approaches with pure motion, and demonstrated the importance of occlusion handling. Our system converges in less than 30 iterations.

Despite the good results, our current system is limited in some ways. Motion is represented as local translational motion. The use of 2D affine or 3D motion models would improve the performance in cases where large rotations or changes of distance take place. Our general approach with its special characteristics however is suited for other motion models. The analysis of ways to handle the additional complexity in the optimization could also be part of future work.

References

[1] Yucel Altunbasak, P. Erhan Eren, and A. Murat Tekalp. Region-based parametric motion segmentation using color information. *Graphical Models and Image Processing*, 60(1):13–23, January 1998.

[2] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.

[3] Patrick Bouthemy and Edouard Francois. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *Intl. Journal of Computer Vision*, 10(2):157–182, April 1993.

[4] D. Cremers and S. Soatto. Motion competition: A variational framework for piecewise parametric motion segmentation. *IJCV*, 62(3):249–265, May 2005.

[5] A. Jepson and M.J. Black. Mixture models for optical flow computation. In *Proc. of CVPR*, pages 760–761, New York, NY, USA, 1993.

[6] S. Khan and M. Shah. Object based segmentation of video using color, motion and spatial information. In *Proc. of CVPR*, volume 2, pages II–746 – II–751, 2001.

[7] Vladimir Kolmogorov, Antonio Criminisi, Andrew Blake, Geoffrey Cross, and Carsten Rother. Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *IEEE Transactions on PAMI*, 28(9):1480–1492, September 2006.

[8] Shang-Hong Lai and Baba C. Vemuri. Reliable and efficient computation of optical flow. *IJCV*, 29(2):87–105, 1998.

[9] D.W. Murray and B.F. Buxton. Scene segmentation from visual motion using global optimization. *IEEE Transactions on PAMI*, 9(2):220–228, March 1987.

[10] C. Stiller. Object-based estimation of dense motion fields. *IEEE Transactions on Image Processing*, 6(2):234–250, February 1997.

[11] L. Torres, M. Kunt, and F. Pereira. Second generation video coding schemes and their role in mpeg-4. In *ECMAST*, pages 799–824, May 1996.

[12] Jun Zhang and G.G. Hanauer. The application of mean field theory to image motion estimation. *IEEE Transactions on Image Processing*, 4(1):19–33, January 1995.

[13] Y.J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.

Author Information:
Dipl.-Inf. Sven Hellbach[1]
Dipl.-Inf Boris Lau[1]
Dr.-Ing. Julian P. Eggert[2]
Prof. Dr. Edgar Körner[2]
Prof. Dr. Horst-Michael Gross[1]
[1]Ilmenau Technical University, Neuroinformatics and Cognitive Robotics Lab, POB 10 05 65, 98684 Ilmenau
[2]Honda Research Institute Europe GmbH, Carl-Legien-Strasse 30, 63073 Offenbach/Main
Phone: +49 3677 69 1306
Fax: +49 3677 69 1665
E-mail: sven.hellbach@tu-ilmenau.de