Segmentation of Range and Intensity Image Sequences by Clustering

Bernd Heisele
M.I.T.
Center of Biological and Computational Learning
Cambridge, MA 02139, USA
heisele@ai.mit.edu

Werner Ritter
DaimlerChrysler Research Center
Image Understanding FT3/AB
Ulm, D-89013, Germany
werner.ritter@daimlerchrysler.com

Abstract

In this contribution we present a method for segmenting temporal sequences of range and intensity images. The paper addresses two problems: Fusion of intensity and range data for image segmentation and visual tracking of segments over time. Our method is based on clustering in a 4D feature space which contains intensity and geometric features. The problem of tracking segments over time is solved by adaptive image sequence clustering. The main idea is to use the cluster centers of the previous image to initialize clustering for the current image. This link between consecutive clustering steps allows to track clusters over time without explicit correspondence analysis. First experiments show that our method can successfully segment and track objects independent of their shapes and motions.

1 Introduction

The task of the proposed algorithm is to segment a temporal sequence of range and intensity images such that a servicing robot can perform visual object tracking in real-time. A laser range camera is mounted on the head of the robot. It generates at each time step a high resolution range and intensity image.

The proposed algorithm addresses two problems I) How to fuse intensity and range information for segmentation and II) how to track segments over time.

I) Basically there are two different approaches to fusing range and intensity data for image segmentation: a) separately segment the range and intensity images and then merge the segmentation results or b) perform the segmentation on the combined range/intensity data. a) When merging two separately segmented images the crucial step is finding corresponding regions in the pair of segmented images. Each segmented image can be represented as graph, where nodes encode the regions and edges the spatial relationships between the regions. The correspondence problem can be seen as a subgraph matching problem which, in general, is NP complete [2]. Once the set of corresponding subgraphs is found, the combined graph is built by choosing one subgraph from each pair of corresponding subgraphs. b) Segmentation of the combined range/intensity data is a more natural approach to the fusion problem. Most of the common image segmentation techniques such as region growing [8], split and merge [4], or clustering [6] can be extended to deal with the combination of range and intensity data. In clustering, data fusion can be achieved by combining different features into a single feature space. The clustering algorithm itself remains unchanged.

II) A main problem in visual tracking is finding corresponding image features (e.g. corner points, edges, regions) in consecutive frames. Usually this problem is treated separately from the problem of feature extraction. In [3] we suggested the combination of feature extraction and tracking. We developed a clustering algorithm for segmentation of color image sequences which implicitly matches corresponding clusters in consecutive images. In [5] the same idea lead to a segmentation method based on region growing. They use the segments of the previous image to initialize the seed points for the region growing algorithm in the current image.

In this paper we will develop a two step clustering technique for segmenting range and intensity image sequences. In the initial step of our algorithm, segments are determined by a divisive clustering algorithm which is applied to the first pair of images. For each new image pair the clusters of the previous pair are adapted iteratively such that links between corresponding clusters are preserved.

The outline of our paper is as follows: In Section 2 we describe the initial clustering of the first image pair. Adaptive image sequence clustering is introduced in Section 3. Our experimental results are presented in Section 4. The paper is summarized in Section 5.

2 Initial Clustering in the combined Position/Intensity Space

At each time step our laser camera delivers a range image and an intensity image as inputs to our segmentation algorithm. Based on the range image and the internal camera parameters we calculate the 3D coordinates for each pixel with respect to the camera coordinate system. The 3D position and intensity features are then combined to a 4D intensity/position feature space where each pixel is described by a vector $\mathbf{f}$, containing its intensity I and its 3D position (X, Y, Z): $\mathbf{f}_{ij} = (w \cdot I_{ij}, X_{ij}, Y_{ij}, Z_{ij})$, where i is the row and j the column of the pixel, w is a weighting factor which determines the relation between the intensity and the position features.

We assume that pixels with similar 3D coordinates and similar intensities belong to the same physical object. Note that this assumption is rather general and does not make any restrictions regarding the shape of the objects. If the assumption holds clustering in the intensity/position feature performs an object-based segmentation since it groups pixels of similar intensity and 3D position. Formally, the task of clustering can be described as to find a number of M prototypes $\mathbf{p}_m$, which minimize the sum of quantization errors: $\sum_{ij} \left(\mathbf{f}_{ij} - \mathbf{p}_{min}\right)^2$, where $\mathbf{p}_{min}$ is the prototype closest to the feature vector $\mathbf{f}_{ij}$ in the intensity/position feature space.

Clustering of the first pair of images in a sequence has to be performed without any prior knowledge about the position of the clusters in the feature space. For this reason, we have chosen a divisive clustering method [6] which does not need to be initialized with cluster positions. The method starts with one cluster covering the whole data space. In each iteration the cluster with the highest variance is split in two by a hyperplane. The hyperplane is determined so that it is perpendicular to the direction of the highest variance and runs through the center of the original cluster. The objective of this heuristic is to minimize the variances of the two new clusters. After the partitioning has been performed, the prototypes are calculated as the centers of the new clusters. Instead of preselecting a maximum number of clusters, one can alternatively define an upper limit for the sum of quantization errors as the stop-criteria for the divisive clustering.

3 Adaptive Image Sequence Clustering

For the first image pair of a sequence a set of prototypes is determined by the divisive clustering described above. For each following image pair the prototypes of the previous pair serve as initialization for the parallel k-means clustering [7]. Parallel k-means clustering consists of two steps per iteration n:

$$\text{Clusters: } C_m(n) = \left\{ \text{pixel } ij \;\middle|\; \left\| \mathbf{f}_{ij} - \mathbf{p}_m(n-1) \right\|^2 \leq \left\| \mathbf{f}_{ij} - \mathbf{p}_l(n-1) \right\|^2 \forall l \right\}, \quad (1)$$

$$\text{Prototypes: } \mathbf{p}_m(n) = \frac{1}{\text{size}\left[C_m(n)\right]} \cdot \sum_{ij \in C_m(n)} \mathbf{f}_{ij}. \quad (2)$$

In the partitioning step each pixel ij, characterized by its feature vector $\mathbf{f}_{ij}$, is assigned to the cluster $C_m(n)$ with the closest prototype $\mathbf{p}_m(n-1)$. After that, the prototypes $\mathbf{p}_m(n)$ are recomputed as the average of the data in their clusters. Each of these two alternating steps reduces the sum of quantization errors until no further changes occur. At this stage a local minimum of the sum of quantization errors is obtained.

Initializing the k-means clustering with prototypes of the previous image implicitly establishes correspondences between clusters in two consecutive frames. If frame-to-frame changes are small, k-means clustering will track scene motion by appropriately shifting the clusters in the 4D feature space. However, initializing with the prototypes of the previous image might not be sufficient to track fast moving objects with large frame-to-frame displacements. For this reason, we added a standard Kalman filter which predicts the location of each prototype in the 4D feature space based on its given trajectory. The predicted prototypes then serve as initialization for the k-means clustering in the next frame. The intensity is assumed to be time invariant and is kept unchanged. Since there can be various types of moving objects in the scene, a rather general kinematic model is chosen in the Kalman filter [1]: The motion along the X-, Y- and Z-axis are assumed to be decoupled, X-, Y- and Z-motions are therefore predicted by separate filters. The motion of the cluster is assumed to have nearby constant velocity. To account for slight changes in the velocity, the time continuous acceleration is modeled as white noise. The discrete state equation for a sampling period T is:

$$\mathbf{s}(k+1) = \mathbf{A}\,\mathbf{s}(k) + \mathbf{w}(k) \quad (3)$$

with

$$\mathbf{A} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix} \quad (4)$$

$$\mathbf{Q} = E\left(\mathbf{w}(k)\mathbf{w}(k)^T\right) = \begin{pmatrix} \frac{1}{3}T^3 & \frac{1}{2}T^2 \\ \frac{1}{2}T^2 & T \end{pmatrix} \sigma_w^2 \quad (5)$$

The measurement equation for the one-dimensional position is:

$$p(k) = \mathbf{C}\,\mathbf{s}(k) + n(k) \quad (6)$$

with

$$\mathbf{C} = \begin{pmatrix} 1 & 0 \end{pmatrix} \quad (7)$$

$$E\left(n^2(k)\right) = \sigma_n^2 \quad (8)$$

The parameters of the filter are the power spectral density of the process noise σ_w^2 and the measurement noise σ_n^2.

4 Experiments

All experiments have been carried out on real intensity/range images taken by a laser range camera which was developed at DaimlerChrysler Aerospace. The camera delivered range data with a relative error of about $\pm 5\%$. The resolution of the range and intensity images were 640×480 pixels. The frame rate of the camera was 7 Hz which lead to large frame-to-frame changes when the camera was mounted on a moving robot or when the camera observed moving objects from a stationary platform. Since our algorithm requires small frame-to-frame changes we had to generate artificial sequences by manually shifting objects between two shots. The next generation of the DaimlerChrysler range camera will have a frame rate of 25 Hz. This will be sufficient to apply our algorithm to natural scenes of moving objects.

We evaluated 3 sequences with about 25 pairs of range and intensity images each. The task of the algorithm was to segment the scene and to track the moving objects. Figures 1 a)–c) show the first intensity images of our 3 test sequences. The dark lines indicate the trajectories along which the objects (box, cylinder, and watering can) have been moved. The frame-to-frame displacements of the moving objects were between 20 and 30 pixels. The benefit of combining intensity and range data is clearly illustrated in sequence 3. At the beginning of sequence 3 (Figures 1 c) and e)) the range data would not be sufficient to separate the watering can from the background. At the end of the sequence (Figures 1 d) and f)) the intensity contrast between the watering can and the occluding box is low, whereas the difference in their range values is significant.

An example of clustering in the combined intensity/position feature space is shown in Figure 2 a). The original intensity of each pixel is replaced by the intensity of the associated cluster center. As expected, large objects are segmented into several tile-like clusters (e.g. dark box in the lower right corner of Figures 2 a)). Most of the object boundaries coincide with boundaries between clusters which indicates that each cluster contains points of only one object. Figures 2 b)–d) show the results of tracking. The dark lines are the trajectories of the clusters which belong to the moving object. The cylinder is segmented into 2, the watering into 5, and the box into 4 clusters. Figures 2 b)and d) show that tracking is robust against partial occlusions.

5 Conclusion

In this paper we proposed a new algorithm for segmenting sequences of range and intensity image pairs. The problem of fusing range and intensity data for image segmentation is solved by clustering in a combined 4D intensity/position feature space. Each image is divided into a given number of clusters by grouping pixels of similar intensity and 3D position. The algorithm unifies segmentation and tracking by initializing the clustering of new images based on clustering results from previous images. In this context, Kalman filters are used to stabilize tracking by predicting dynamic changes in cluster positions. First experiments have been carried out on real range/intensity images taken by a laser camera. The method successfully segmented and tracked moving objects of various shapes.

References

[1] Y. Bar-Shalom and X.-R. Li. *Estimation and tracking: Principles, techniques, and software*. Artech House, Boston, 1993.

[2] N. Deo. *Graph theory with applications to engineering and computer science*. Prentice-Hall, 1974.

[3] B. Heisele, U. Kressel, and W. Ritter. Tracking non-rigid, moving objects based on color cluster flow. In *Proc. Computer Vision and Pattern Recognition*, pages 253–257, San Juan, 1997.

[4] S. L. Horowitz and T. Pavlidis. Picture segmentation by a tree traversal algorithm. *Journal of the ACM*, 23:368–388, 1976.

[5] X. Jiang, S. Hofer, T. Stahs, I. Ahrns, and H. Bunke. Extraction and tracking of surfaces in range image sequences. In *Second International Conference on 3D Digital Imaging and Modeling*, Ottawa, Canada, 1999.

[6] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[7] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkley Symposium on Mathematics Statistics and Probability*, pages 281–297, 1967.

[8] S. W. Zucker. Region growing: Childhood and adolescence. *Computer Graphics and Image Processing*, 5:382–399, 1976.

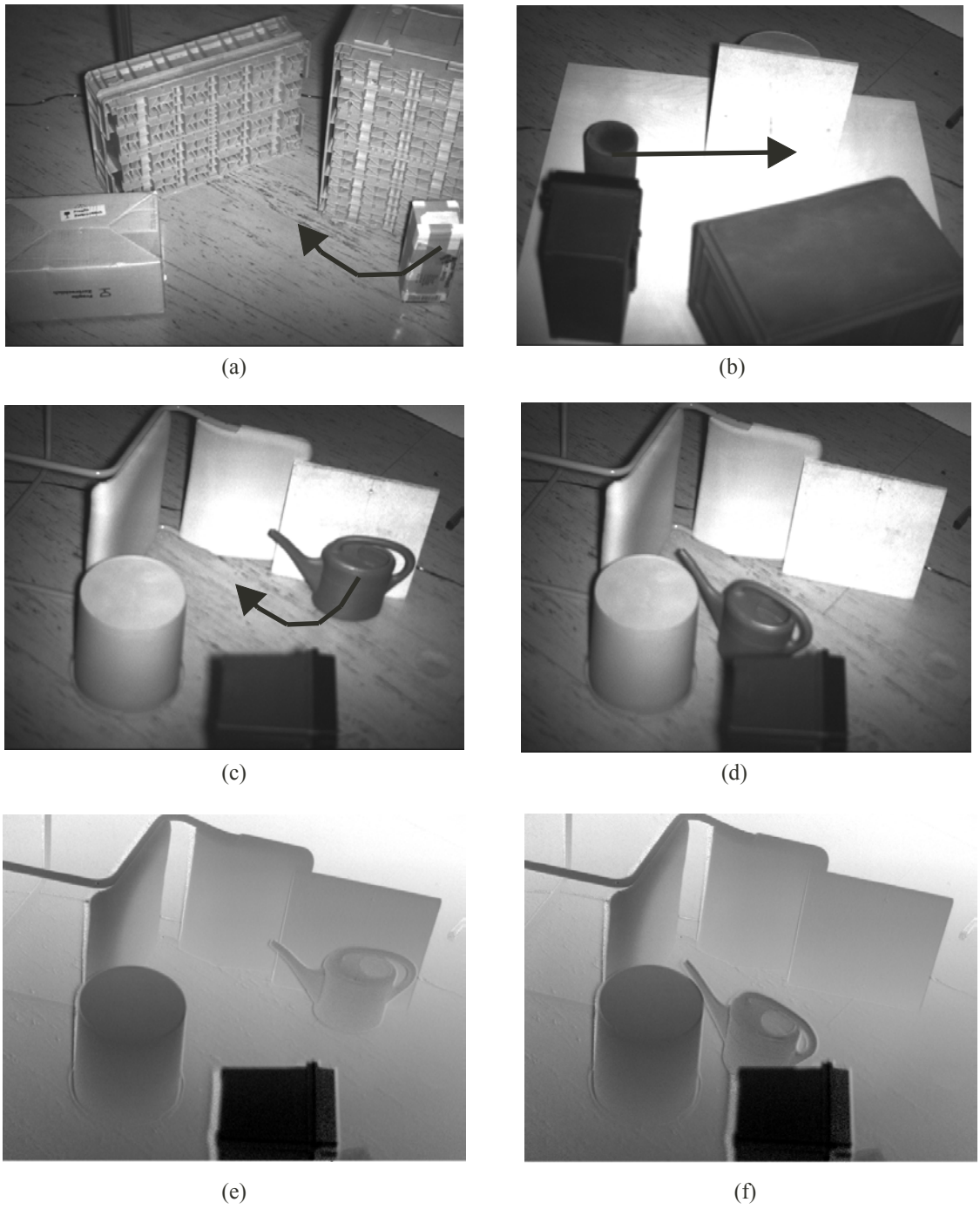

Figure 1. a)–d): Intensity images of our 3 test sequences. The dark lines indicate the trajectories along which the objects have been moved. e)–f): First and last range image of the third test sequence.

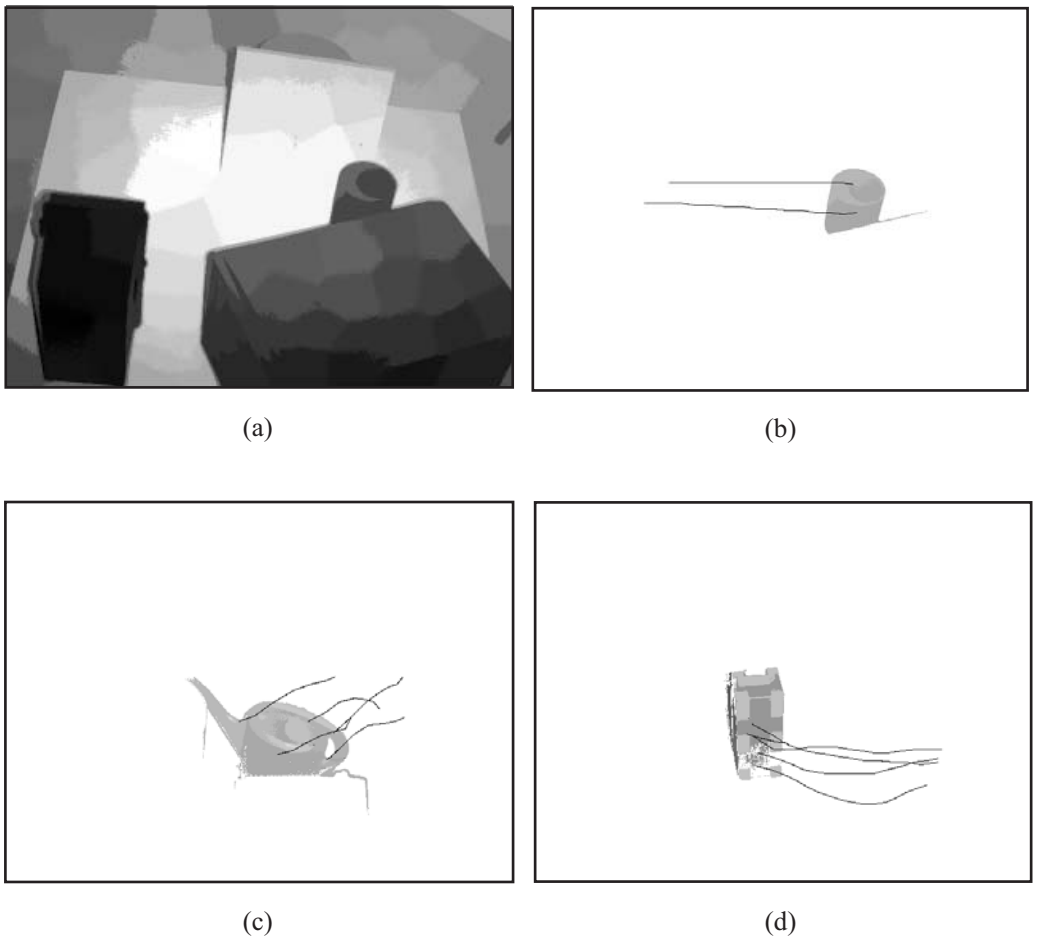

Figure 2. a): Results of the intensity/position clustering. The original intensity of each pixel is replaced by the intensity of the associated cluster center. b)–d): Results of tracking. Black lines represent the trajectories of the cluster centers.