What Went Where

Josh Wills, Sameer Agarwal and Serge Belongie

Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA
{josh,sagarwal,sjb} @ cs.ucsd.edu

Abstract

We present a novel framework for motion segmentation that combines the concepts of layer-based methods and feature-based motion estimation. We estimate the initial correspondences by comparing vectors of filter outputs at interest points, from which we compute candidate scene relations via random sampling of minimal subsets of correspondences. We achieve a dense, piecewise smooth assignment of pixels to motion layers using a fast approximate graph-cut algorithm based on a Markov random field formulation. We demonstrate our approach on image pairs containing large inter-frame motion and partial occlusion. The approach is efficient and it successfully segments scenes with inter-frame disparities previously beyond the scope of layer-based motion segmentation methods.

1. Introduction

The problem of motion segmentation consists of (1) finding groups of pixels in two or more frames that move together, and (2) recovering the motion fields associated with each group. Motion segmentation has wide applicability in areas such as video coding, content-based video retrieval, and mosaicking. In its full generality, the problem cannot be solved since infinitely many constituent motions can explain the changes from one frame to another. Fortunately, in real scenes the problem is simplified by the observation that objects are usually composed of spatially contiguous regions and the number of independent motions is significantly smaller than the number of pixels. Operating under these assumptions, we propose a new motion segmentation algorithm for scenes containing objects with large inter-frame motion. The algorithm operates in two stages, starting with robust estimation of the underlying motion fields and concluding with dense assignment of pixels to motion fields. This work is the first dense motion segmentation method to operationalize the layer-based formulation for multiple discrete motions.

The structure of the paper is as follows. We will begin in section 2 with an overview of related work. In section 3, we detail the components of our approach. We discuss experimental results in section 4. The paper concludes with discussion in section 5.

2. Related Work

Early approaches to motion segmentation were based on estimating dense optical flow. The optical flow field was assumed to be piecewise smooth to account for discontinuities due to occlusion and object boundaries. Wang & Adelson introduced the idea of decomposing the image sequence into multiple overlapping layers, where each layer is a smooth motion field [13].

Optical flow based methods are limited in their ability to handle large inter-frame motion or objects with overlapping motion fields. Coarse-to-fine methods are able to solve the problem of large motion to a certain extent but the degree of sub-sampling required to make the motion differential places an upper bound on the maximum allowable motion between two frames and limits it to about 15% of the dimensions of the image. Also in cases where the order of objects along any line in the scene is reversed and their motion fields overlap, the coarse to fine processing ends up blurring the two motions into a single motion before optical flow can be calculated.

In this paper we are interested in the case of discrete motion, i.e. where optical flow based methods break down. Most closely related to our work is that of Torr [11]. Torr uses sparse correspondences obtained by running a feature detector and matching them using normalized cross correlation. He then processes the correspondences in a RANSAC framework to sequentially cover the the set of motions in the scene. Each iteration of his algorithm finds the dominant motion model that best explains the data and is simplest according to a complexity measure. The set of models and the associated correspondences are then used as the initial guess for the estimation of a mixture model using the Expectation Maximization (EM) algorithm. Spurious models are pruned and the resulting segmentation is smoothed

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

using morphological operations.

In a more recent work [12], the authors extend the model to 3D layers in which points in the layer have an associated disparity. This allows for scenes in which the planarity assumption is violated and/or a significant amount of parallax is present. The pixel correspondences are found using a multiscale differential optical flow algorithm, from which the layers are estimated in a Bayesian framework using EM. Piecewise smoothness is ensured by using a Markov random field prior.

Neither of the above works demonstrate the ability to perform dense motion segmentation on a pair of images with large inter-frame motion. In both of the above works the grouping is performed in a Bayesian framework. While the formulation is optimal and strong results can be proved about the optimality of the Maximum Likelihood solution, actually solving for it is an extremely hard non-linear optimization problem. The use of EM only guarantees a locally optimal solution and says nothing about the quality of the solution. As the authors point out, the key to getting a good segmentation using their algorithm is to start with a good guess of the solution and they devote a significant amount of effort to finding such a guess. However it is not clear from their result how much the EM algorithm improves upon their initial solution.

3. Our Approach

Our approach is based on a two stage process the first of which is responsible for motion field estimation and the second of which is responsible for motion layer assignment. As a preliminary step we detect interest points in the two images and match them by comparing filter responses. We then use a RANSAC based procedure for detecting the motion fields relating the frames. Based on the detected motion fields, the correspondences detected in the first stage are partitioned into groups corresponding to a single motion field and the resulting motion fields are re-estimated. Finally, we use a fast approximate graph cut based method to densely assign pixels to their respective motion fields. We now describe each of these steps in detail.

3.1. Interest point detection and matching

Many pixels in real images are redundant so it is beneficial to find a set of points that reduce some of this redundancy. To achieve this, we detect interest points using the Förstner operator [4]. To describe each interest point, we apply a set of 76 filters (3 scales and 12 orientations with even and odd phase and an elongation ratio of 3:1, plus 4 spot filters) to each image. The filters, which are at most 31×31 pixels in size, are evenly spaced in orientation at intervals of 15° and the changes in scale are half octave. For each of the scales and orientations, there is a quadrature pair of derivative-of-Gaussian filters corresponding to edge and bar-detectors respectively, as in [7, 5].

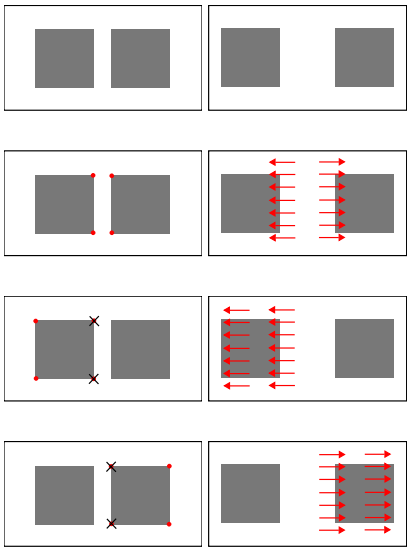

Figure 1: Phantom motion fields. (Row 1) Scene that consists of two squares translating away from each other. (Row 2) Under an affine model, triplets of points that span the two squares will propose a global stretching motion. This motion is likely to have many inliers since all points on the inner edges of the squares will fit this motion exactly. If we then delete all points that agree with this transformation, we will be unable to detect the true motions of the squares in the scene (Rows 3 & 4).

To obtain some degree of rotational invariance, the filter response vectors may be reordered so that the order of orientations is cyclically shifted. This is equivalent to filtering a rotated version of the image patch that is within the support of the filter. We perform three such rotations in each direction to obtain rotational invariance up to $\pm 45^\circ$.

We find correspondences by comparing filter response vectors using the L_1 distance. We compare each interest point in the first image to those in the second image and assign correspondence between points with minimal error. Since matching is difficult for image pairs with large inter-frame disparity, the remainder of our approach must take into account that the estimated correspondences can be extremely noisy.

3.2. Estimating Motion Fields

Robust estimation methods such as RANSAC [3] have been shown to provide very good results in the presence of noise when estimating a single, global transformation between images. Why can't we simply apply these methods to multi-

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

ple motions directly? It turns out that this is not as straightforward as one might imagine. Methods in this vein work by iteratively repeating the estimation process where each time a dominant motion is detected, all correspondences that are deemed inliers for this motion are removed [11].

There are a number of issues that need to be addressed before RANSAC can be used for the purpose of detecting and estimating multiple motions. The first issue is that combinations of correspondences – not individual correspondences – are what promote a given transformation. Thus when "phantom motion fields" are present, i.e. transformations arising from the relative motion between two or more objects, it is possible that the deletion of correspondences could prevent the detection of the true independent motions; see Figure 1. Our approach does not perform sequential deletion of correspondences and thus circumvents this problem.

Another consideration arises from the fact that the RANSAC estimation procedure is based on correspondences between interest points in the two images. This makes the procedure biased towards texture rich regions, which have a large number of interest points associate with them, and against small objects in the scene, which in turn have a small number of interest points. In the case where there is only one global transformation relating the two images, this bias does not pose a problem. However it becomes apparent when searching for multiple independent motions. To correct for this bias we introduce "perturbed interest points" and a method for feature crowdedness compensation.

3.2.1. Perturbed Interest Points

If an object is only represented by a small number of interest points, it is unlikely that many samples will fall entirely within the object. One approach for boosting the effect of correct correspondences without boosting that of the incorrect correspondences is to appeal to the idea of a stable system. According to the principle of perturbation, a stable system will remain at or near equilibrium even as it is slightly modified. The same holds true for stable matches. To take advantage of this principle, we dilate the interest points to be disks with a radius of r_p, where each pixel in the disk is added to the list of interest points. This allows the correct matches to get support from the points surrounding a given feature while incorrect matches will tend to have almost random matches estimated for their immediate neighbors, which will not likely contribute to a widely-supported warp. In this way, while the density around a valid motion is increased, we do not see the same increase in the case of an invalid motion; see Figure 2.

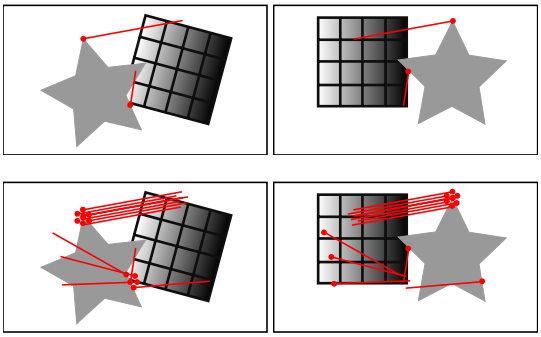

Figure 2: Perturbed Interest Points. Correspondences are represented by point-line pairs where the point specifies an interest point in the image and the line segment ends at the location of the corresponding point in the other image. (1) We see one correct correspondence and one incorrect correspondence that is the result of an occlusion junction forming a white wedge. (2) The points around the correct point have matches that are near the corresponding point, but the points around the incorrect correspondence do not.

3.2.2. Feature Crowdedness

Textured regions often have significant representation in the set of interest points. This means that a highly textured object will have a much larger representation in the set of interest points than an object of the same size with less texture. To mitigate this effect, we bias the sampling. We calculate a measure of crowdedness for each interest point and the probability of choosing a given point is inversely proportional to this crowdedness score. The crowdedness score is the number of interest points that fall into a disk of radius r_c.

3.2.3. Partitioning and Motion Estimation

Having perturbed the interest points and established a sampling distribution on them, we are now in a position to detect the motions present in the frames. We do so using a two step variant of RANSAC, where multiple independent motions are explicitly handled, as duplicate transformations are detected and pruned in a greedy manner. The first step provides a rough partitioning of the set of correspondences (motion identification) and the second takes this partitioning and estimates the motion of each group.

First, a set of planar warps is estimated by a round of standard RANSAC and inlier counts (using an inlier threshold of τ) are recorded for each transformation. In our case, we use planar homography which requires 4 correspondences to estimate, however similarity or affinity may be used (requiring 2 and 3 correspondences, respectively). The estimated list of transformations is then sorted by inlier count and we keep the first n_t transformations, where n_t is

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

some large number (e.g. 300).

We expect that the motions in the scene will likely be detected multiple times and we would like to detect these duplicate transformations. Comparing transformations in the space of parameters is difficult for all but the simplest of transformations, so we compare transformations by comparing the set of inliers associated with each transformation. If there is a large overlap in the set of inliers (more than 75%) the transformation with the larger set of inliers is kept and the other is pruned.

Now that we have our partitioning of the set of correspondences, we would like to estimate the planar motion represented in each group. This is done with a second round of RANSAC on each group with only 100 iterations. This round has a tighter threshold to find a better estimate. We then prune duplicate warps a second time to account for slightly different inlier sets that converged to the same transformation during the second round of RANSAC with the tighter threshold.

The result of this stage is a set of proposed transformations and we are now faced with the problem of assigning each pixel to a candidate motion field.

3.3. Layer Assignment

The problem of assigning each pixel to a candidate motion field can be formulated as finding a function $l : I \rightarrow \{1, \ldots, m\}$, that maps each pixel to an integer in the range $1, \ldots, m$, where m is the total number of of motion fields, such that the reconstruction error

$$\sum_i [I(i) - I'(M(l(i), i))]^2$$

is minimized. Here $M(p, q)$ returns the position of pixel q under the influence of the motion field p.

A naïve approach to solving this problem is to use a greedy algorithm that assigns each pixel the motion field for which it has the least reconstruction error, i.e.

$$l(i) = \underset{1 \leq p \leq m}{\operatorname{argmin}} [I(i) - I'(M(p, i))]^2 \qquad (1)$$

The biggest disadvantage of this method as can be seen in Figure 3 is that for flat regions it can produce unstable labellings, in that neighboring pixels that have the same brightness and are part of the same moving object can get assigned to different warps. What we would like instead is to have a labelling that is piecewise constant with the occasional discontinuity to account for genuine changes in motion fields.

The most common way this problem is solved (see e.g. [14]) is by imposing a smoothness prior over the set of solutions, i.e. an ordering that prefers piecewise constant labellings over highly unstable ones. It is important that the prior be sensitive to true discontinuites present in the image. In [1], for example, Boykov, Veksler and Zabih have shown that discontinuity preserving smoothing can be performed by adding a penalty of the following form to the objective function

$$\sum_i \sum_{j \in \mathcal{N}(i)} s_{ij} [1 - \delta_{l(i)l(j)}]$$

Given a measure of similarity s_{ij} between pixels i and j, it penalizes pixel pairs that have been assigned different labels. The penalty should only be applicable for pixels that are near each other. Hence the second sum is over a fixed neighborhood $\mathcal{N}(i)$. The final objective function we minimize is

$$\sum_i [I(i) - I'(M(l(i), i))]^2 + \lambda \sum_i \sum_{j \in \mathcal{N}(i)} s_{ij} [1 - \delta_{l(i)l(j)}]$$

where λ is the tradeoff between the data and the smoothness prior.

An optimization problem of this form is known as a *Generalized Potts model* which in turn is special case of a class of problems known as metric labelling problems. Kleinberg & Tardos demonstrate that the metric labelling problems corresponds to finding the maximum *a posteriori* labelling of a class of Markov random field [8]. The problem is known to be NP-complete, and the best one can hope for in polynomial time is an approximation.

Recently Boykov, Veksler and Zabih have developed a polynomial time algorithm that finds a solution with error at most two times that of the optimal solution [2]. Each iteration of the algorithm constructs a graph and finds a new labelling of the pixels corresponding to the minimum cut partition in the graph. The algorithm is deterministic and guaranteed to terminate in $O(m)$ iterations.

Besides the motion fields and the image pair, the algorithm takes as input a similarity measure s_{ij} between every pair of pixels i, j within a fixed distance of one another and two parameters, k the size of the neighborhood around each pixel, and λ the tradeoff between the data and the smoothness term. We use a Gaussian weighted measure of the

Figure 3: Example of naïve pixel assignment as in Equation 1 for the second motion layer in Figure 5. Notice there are many pixels that are erratically assigned. This is why smoothing is needed.

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

squared difference between the intensities of pixels i and j,

$$s_{ij} = \exp\left[-\frac{d(i,j)^2}{2k^2} - (I(i) - I(j))^2\right]$$

where $d(i,j)$ is the distance between pixel i and pixel j.

We run the above algorithm twice, once to assign the pixels in the image I to the forward motion field and again to assign the pixels in image I' to the inverse motion fields relating I' and I. If a point in the scene occurs in both frames, we expect that its position and appearance will be related as:

$$\begin{aligned} M(l(p),p) &= p' \\ M(l'(p'),p') &= p \\ I(M(l(p),p)) &= I'(p) \end{aligned}$$

Here, the unprimed symbols refer to image I and the primed symbols refer to image I'. Assuming that the appearance of the object remains the same across the images, the final assignment is obtained by intersecting the forward and backward assignments.

```
1.  Detect interest points in I
2.  Perturb each interest point
3.  Find the matching points in I'
4.  For i = 1:N_s
       Pick tuples of correspondences
       Estimate the warp
       Store inlier count
5.  Prune the list of warps
6.  Refine each warp using its inliers
7.  Perform dense pixel assignment
```

Figure 4: Algorithm Summary

4. Experimental Results

We now illustrate our algorithm, which is summarized in Figure 4, on several pairs of images containing objects undergoing independent motions. We performed all of the experiments on grayscale images with the same parameters[1].

Our first example is shown in Figure 5. In this figure we show the two images, I and I', and the assignments for each pixel to a motion layer (one of the three detected motion fields). The rows represent the different motion fields and the columns represent the portions of each image that are assigned to a given motion layer. The motions are made explicit in that the pixel support from frame to frame is related exactly by a planar homography. Notice that the portions of the background and the dumpsters that were visible in both frames were segmented correctly, as was the man. This example shows that in the presence of occlusion and when visual correspondence is difficult (i.e. matching the dumpsters correctly), our method provides good segmentation. Another thing to note is that the motion of the man is only approximately planar.

[1] $N_s = 10^4, n_t = 300, r_p = 2, r_c = 25, \tau = 10, k = 2, \lambda = .285$

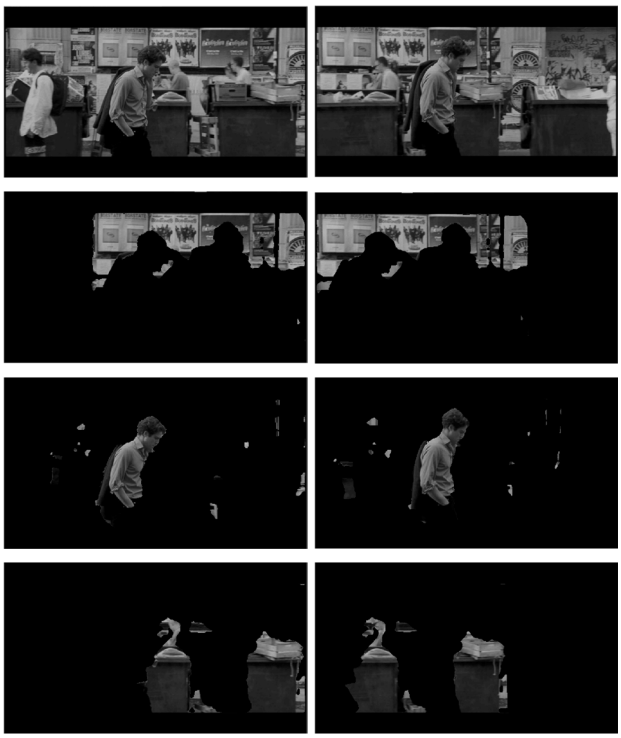

Figure 5: Notting Hill sequence. (1) Original image pair of size 311×552, (2-4) Pixels assigned to warp layers 1-3 in I and I'.

Figure 6 shows a scene consisting of two fish swimming past a fairly complicated reef scene. The segmentation is shown as in Figure 5 and we see that three motions were detected, one for the background and one for each of the two fish. In this scene, the fish are small, feature-impoverished objects in front of a large feature-rich background, thus making the identification of the motion of the fish difficult. In fact, when this example was run without using the perturbed interest points, we were unable to recover the motion of either of the fish.

Figure 7 shows two frames from a sequence that has been a benchmark for motion segmentation approaches for some time. Previously, only optical flow-based techniques were able to get good motion segmentation results for this scene, however producing a segmentation of the motion between the two frames shown (1 and 30) would require using all (or at least most) of the intermediate frames. Here the only input to the system was the frames shown. Notice that the portions of the house and the garden that were visible in both frames were segmented accurately as was the tree. This example shows the discriminative power of our filterbank as we were unable to detect the motion field correctly using

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

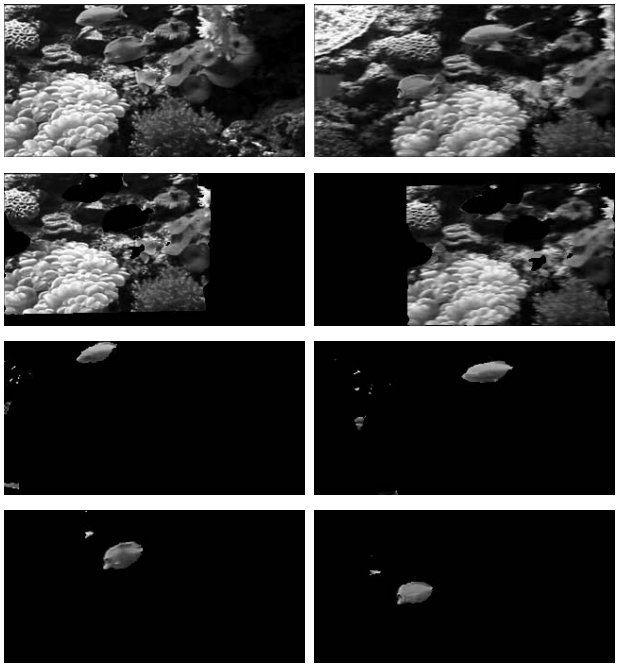

Figure 6: Fish sequence. (1) Original image pair of size 162×319, (2-4) Pixels assigned to warp layers 1-3 in I and I'.

correspondences found using the standard technique of normalized cross correlation.

In Figure 8, a moving car passes behind a tree as the camera pans. Here, only two motion layers were recovered and they correspond to the static background and to the car. Since a camera rotating around its optical axis produces no parallax for a static scene, the tree is in the same motion layer as the fence in the background, whereas the motion of the car requires its own layer. The slight rotation in depth of the car does not present a problem here.

As a final experiment, we demonstrate an application of our algorithm to the problem of video object deletion in the spirit of [6, 13]; see Figure 9. The idea of using motion segmentation information to fill in occluded regions is not new, however previous approaches require a high frame rate to ensure that inter-frame disparities are small enough for differential optical flow to work properly. Here the interframe disparities are as much as a third of the image width.

5. Discussion

In this paper we have presented a new method for performing dense motion segmentation in the presence of large inter-frame motion. Like any system, our system is limited by the assumptions it makes. We make three assumptions about the scenes: 1) identifiability, 2) constant appearance, 3) planar motion.

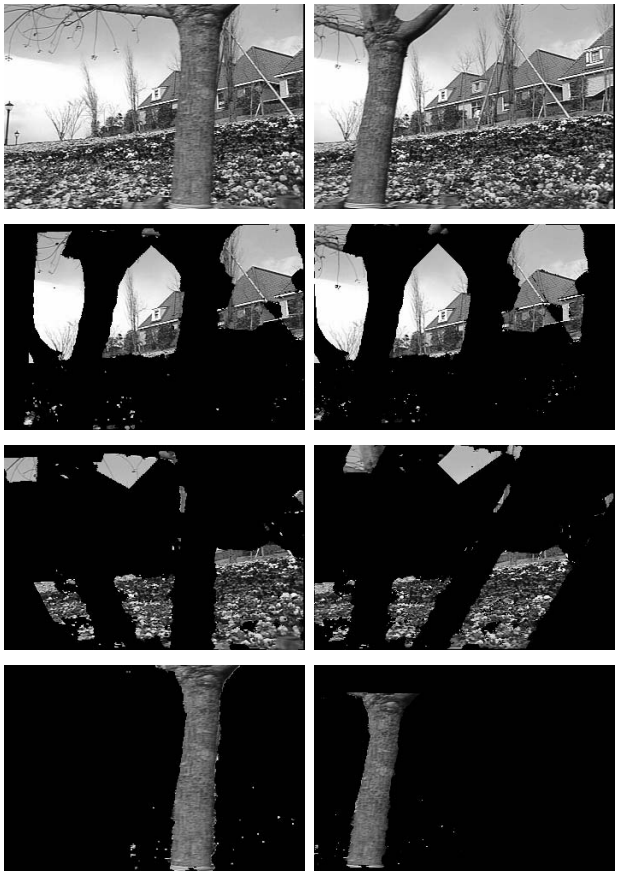

Figure 7: Flower Garden sequence. (1) Original image pair of size 240×360, (2-4) Pixels assigned to warp layers 1-3 in I and I'.

A system is identifiable if its internal parameters can be estimated given the data. In the case of motion segmentation it implies that given a pair of images it is possible to recover the underlying motion. The minimal requirement under our chosen motion model is that each object present in the two scenes should be uniquely identifiable. Consider Figure 10; in this display, several motions can relate the two frames, and unless we make additional assumptions about the underlying problem, it is ill posed and cannot be solved.

Our second assumption is that the appearance of an object across the two frames remains the same. While we do not believe that this assumption can be done away with completely, it can be relaxed. Our feature extraction, description, and matching is based on a fixed set of filters. This gives us a limited degree of rotation and scale invariance. We believe that the matching stage of our algorithm can benefit from the work on affine invariant feature point description [10] and feature matching algorithms based on spatial propagation of good matches [9].

The third assumption of a planar motion model is not a serious limitation, and preliminary experiments show that

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

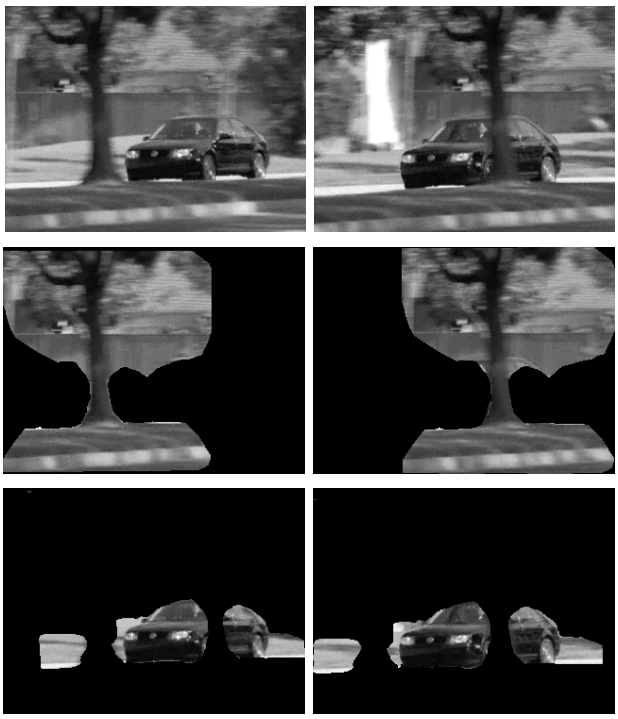

Figure 8: VW sequence. (1) Original image pair of size 240×320, (2-3) Pixels assigned to warp layers 1-2 in I and I'.

extensions to non-planar motion are straightforward.

Finally, in some of the examples we can see that while the segments closely match the individual objects in the scene, some of the background bleeds into each layer. Motion is just one of several cues used by the human vision system in perceptual grouping and we cannot expect a system based purely on the cues of motion and brightness to be able to do the job. Incorporation of the various Gestalt cues and priors on object appearance will be the subject of future research.

Figure 10: Ternus Display. The motion of the dots is ambiguous; additional assumptions are needed to recover their true motion.

Acknowledgments

We would like to thank Charless Fowlkes, Henrik Wann Jensen, David Kriegman, and David Martin for helpful discussions. We would also like to thank Yuri Boykov and Olga Veksler for helpful comments and min-cut code.

This work was partially supported under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Approximate energy minimization with discontinuities. In *IEEE International Workshop on Energy Minimization Methods in Computer Vision*, pages 205–220, 1999.

[2] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239, 2001.

[3] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, vol. 24:381–95, 1981.

[4] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Intercommission Conference on Fast Processing of Photogrammetric Data*, pages 281–305, Interlaken, Switzerland, 1987.

[5] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(9):891–906, September 1991.

[6] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, December 1993.

[7] D. Jones and J. Malik. Computational framework to determining stereo correspondence from a set of linear spatial filters. *Image and Vision Computing*, 10(10):699–708, Dec. 1992.

[8] J. Kleinberg and E. Tardos. Approximate algorithms for classification problems with pairwise relationships: Metric labelling and markov random fields. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, 1999.

[9] M. Lhuillier and L. Quan. Match propagation for image-based modeling and rendering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1140–1146, 2002.

[10] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142. Springer, 2002. Copenhagen.

[11] P. H. S. Torr. Geometric motion segmentation and model selection. In J. Lasenby, A. Zisserman, R. Cipolla, and H. Longuet-Higgins, editors, *Philosophical Transactions of the Royal Society A*, pages 1321–1340. Roy Soc, 1998.

[12] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated Bayesian approach to layer extraction from image sequences. In *Seventh International Conference on Computer Vision*, volume 2, pages 983–991, 1999.

[13] J. Wang and E. H. Adelson. Layered representation for motion analysis. In *Proc Conf. Computer Vision and Pattern Recognition*, pages 361–366, 1993.

[14] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, pages 520–526, 1997.

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

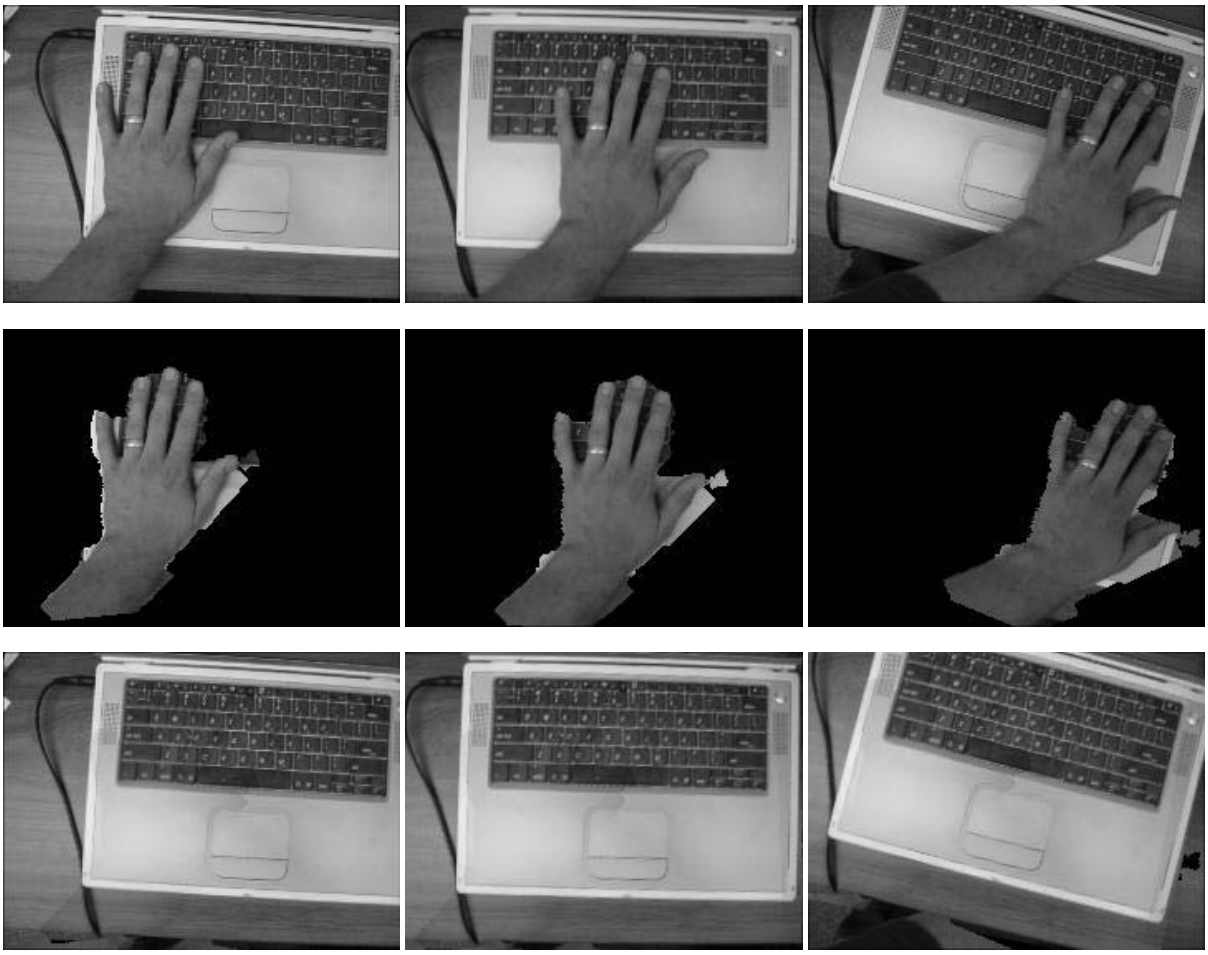

Figure 9: Illustration of video object deletion. (1) Original frames of size 180×240. (2) Segmented layer corresponding to the motion of the hand. (3) Reconstruction without the hand layer using the recovered motion of the keyboard. Note that no additional frames beyond the three shown were used as input.

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE