M.I.T. Media Laboratory Vision and Modeling Group, Technical Report No. 221, April 1993
Appears in *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*,
pp. 361–366, New York, June 1993.

Layered Representation for Motion Analysis

John Y. A. Wang
The MIT Media Laboratory
Dept. Elec. Eng. and Comp. Sci.
Massachusetts Institute of Technology
Cambridge, MA 02139

Edward H. Adelson
The MIT Media Laboratory
Dept. Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Standard approaches to motion analysis assume that the optic flow is smooth; such techniques have trouble dealing with occlusion boundaries. The most popular solution is to allow discontinuities in the flow field, imposing the smoothness constraint in a piecewise fashion. But there is a sense in which the discontinuities in flow are artifactual, resulting from the attempt to capture the motion of multiple overlapping objects in a single flow field. Instead we can decompose the image sequence into a set of overlapping layers, where each layer's motion is described by a smooth flow field. The discontinuities in the description are then attributed to object opacities rather than to the flow itself, mirroring the structure of the scene. We have devised a set of techniques for segmenting images into coherently moving regions using affine motion analysis and clustering techniques. We are able to decompose an image into a set of layers along with information about occlusion and depth ordering. We have applied the techniques to the "flower garden" sequence. We can analyze the scene into four layers, and then represent the entire 30-frame sequence with a single image of each layer, along with associated motion parameters.

1 Introduction

Occlusions represent one of the difficult problems in motion analysis. Smoothing is necessary in order to derive reliable flow fields, but when smoothing occurs across boundaries the result is a flow field that is simply incorrect. Various techniques have been devised to allow for motion discontinuities but none are entirely satisfactory. In addition, transparency due to various sources (including motion blur) can make it meaningless to assign a single motion vector to a single point. It is helpful to reconsider this problem from a different point of view.

Consider an image that is formed by one opaque object moving in front of a background. In Figure 1, this is illustrated with a moving hand in front of a stationary checkerboard. The first row shows the objects that compose the scene; the second row shows the image sequence that will result. An animation system – whether traditional cel animation or modern digital compositing – can generate this sequence by starting with an image of the background, an image of the hand, an opacity map (known as a "matte" or an "alpha channel") for the hand, motion fields for the hand and the background, and finally the rules of image formation.

The resulting image sequence will pose challenges for standard motion analysis because of the occlusion boundaries. But in principle we should be able to retrieve the same simple description of the sequence that the animator used in generating it: an opaque hand moving smoothly in front of a background. The desired description for the hand is shown in the third row of Figure 1; it involves an intensity map, an opacity map, and a warp map. The background (not shown) would also be extracted. Having accomplished this decomposition we could transmit the information very efficiently and could then resynthesize the original sequence, as shown in the bottom row. In addition, the description could be an important step on the way to a meaningful object-based description of the scene, rather than a mere description of a flow field.

Adelson [1] has described a general framework for "layered image representation," in which image sequences are decomposed into a set of layers ordered in depth along with associated maps defining their motions, opacities, and intensities. Given such a description, it is straightforward to synthesize the image sequence using standard techniques of warping and compositing. The challenge is to achieve the description starting with an image sequence from a natural scene. In other words: rendering is easy, but vision

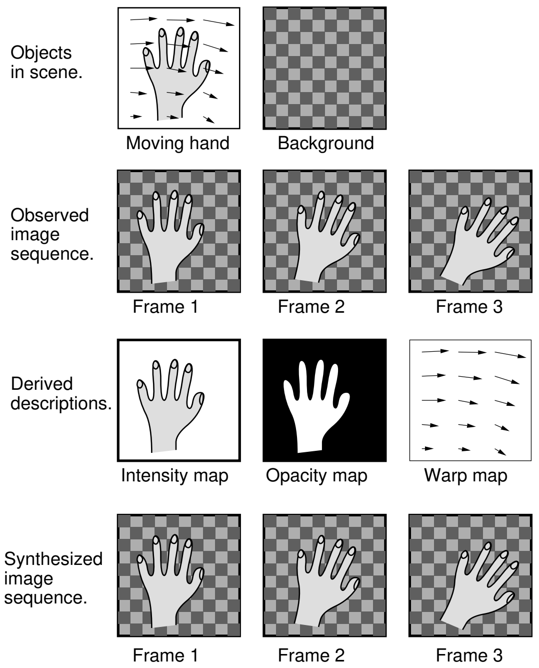

Figure 1: This figures show the decomposition of an image sequence consisting of a hand moving in front of a checkerboard background. The conventional method of representing motion is by a dense motion field with motion discontinuities at object boundary. The layered representation describes these objects with smooth motions, and discontinuities in opacity. The apparent motion discontinuities result when the layers are composited according to the occlusion relationship between objects.

is difficult, as usual. In this paper we describe some techniques that are helpful in accomplishing the vision side of the procedure.

2 Image analysis

Analysis of the scene into the layered representation requires grouping the points in the image into multiple regions where each region undergoes a smooth motion. However, multiple motion estimation and segmentation is a difficult problem that involves a simultaneous estimation of the object boundary and motion. Without the knowledge of the object boundaries, motion estimation will incorrectly apply the image constraints across multiple objects. Likewise, object boundaries are difficult to determine without some estimation of motion.

Recent works by [7, 2, 9] have shown that the affine motion model provides a good approximation of 3-D moving objects. Since the motion model used in the analysis will determine the descriptiveness the representation, we use the affine motion model in our layered representation to describe a wide range of motions commonly encountered in image sequences. These motions include translation, rotation, zoom, and shear. Affine motion is parameterized by six parameters as follows:

$$V_x(x,y) = a_{x0} + a_{x1}\ x + a_{x2}\ y; \tag{1}$$

$$V_y(x,y) = a_{y0} + a_{y1}\ x + a_{y2}\ y \tag{2}$$

where at each point (x, y), $V_x(x, y)$ and $V_y(x, y)$ are the x and y components of velocity respectively, and the a_k's are the affine motion parameters.

3 Implementation

Typical methods in multiple affine motion estimation use an iterative motion estimation techniques to detect multiple affine motion regions in the scene. At each iteration, these methods assume that a dominant motion region can be detected and eliminated from subsequent analysis. Estimation of these regions involve global estimation using a single motion model, and thus, often result in accumulating data from multiple objects.

Our implementation of multiple motion estimation is similar to robust techniques presented by [3, 4, 5]. We use a gradual migration from a local motion representation to a global object motion representation. By performing optic flow estimation follow by affine estimation instead of a direct global affine motion estimation, we can minimize the problems of multiple objects within our analysis region. The layer's image, opacity map are obtained by integrating the motion and regions over time. Our analysis of an image sequence into layers consists of three stages: 1) local motion estimation; 2) motion-based segmentation and; 3) object image recovery.

3.1 Motion segmentation

Our motion segmentation algorithm is illustrated in Figure 2. The segmentation algorithm is is divided into two primary steps: 1) local motion estimation and, 2) affine motion segmentation. Multiple affine motions are estimated within subregions of the image and coherent motion regions are determined based on the estimated affine models. By iteratively updating the affine models and the regions, this architecture

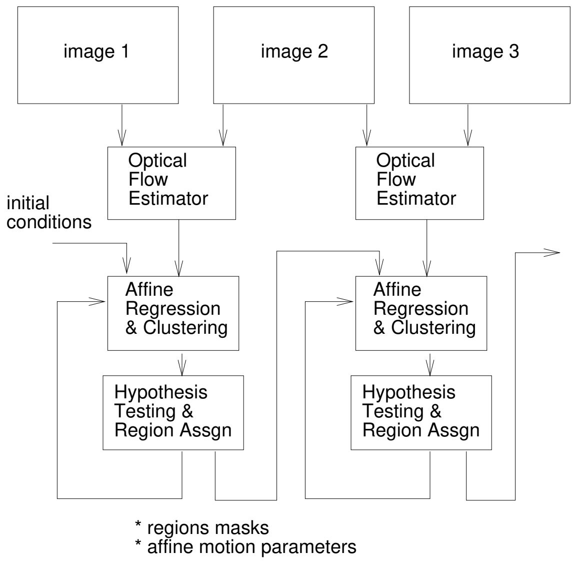

Figure 2: This figures shows the technique used in motion segmentation. Affine motion models are determined by regression on the dense motion fields and the regions are assigned to minimizes the error between the motion expected by the models and the estimated dense motion.

minimizes the problem of intergrating data across object boundaries.

Our local motion estimation is obtained with a multi-scale coarse-to-fine algorithm based on a gradient approach described by [8]. Since only one motion is visible at any point when dealing with opaque objects, the single motion model assumed in the optic flow estimation is acceptable. The multi-scale implementation allows for estimation of large motions. When analyzing scene exhibiting transparent phenomena, the motion estimation technique described by Shizawa and Mase [10] may be suitable.

Motion segmentation is obtained by iteratively refining the estimates of affine motions and the corresponding regions. We estimate the affine parameters within each subregion of the image by standard regression techniques on local motion field. This estimation can be seen as a plane fitting algorithm in the velocity space since the affine model is a linear model of local motion. The regression is applied separately on each velocity component since the components are independent. If we let $H_i = [H_{y_i} \quad H_{x_i}]$ be the i^{th} hypothesis vector in the affine parameter space with components ${H_{x_i}}^T = [a_{x0i} \quad a_{x1i} \quad a_{x2i}]$ and ${H_{y_i}}^T = [a_{y0_i} \quad a_{y1_i} \quad a_{y2_i}]$ corresponding to the x and y components , and $\phi^T = [1 \quad x \quad y]$ be the regressor, then the affine equations 1 and 2 become:

$$V_x(x,y) = \phi^T H_{x_i} \tag{3}$$

$$V_y(x,y) = \phi^T H_{y_i} \tag{4}$$

and a linear least squares estimate of H_i for an given local motion field is as follows:

$$[H_{y_i} \quad H_{x_i}] = [\sum_{P_i} \phi \ \phi^T]^{-1} \sum_{P_i} (\phi [V_y(x,y) \quad V_x(x,y)]) \tag{5}$$

The summation is taken over P_i corresponding to the i^{th} subregion in the image.

We avoid estimating motion across object boundaries by initially using small arbitrary subregions within the image to obtain a set of hypotheses of likely affine motions exhibited in the image. Many of these hypotheses will be incorrect because these initial subregions may contain object boundaries. We identify these hypotheses by their large residual error and eliminate them from our analysis.

However, motion estimates from patches that cover the same object will have similar parameters. These are grouped in the affine motion parameter space with a k-means clustering algorithm described in [11]. In the clustering process, we derive a representative model for each group of similar models. The model clustering produces a set of likely affine motion models that are exhibited by objects in the scene.

Next, we use hypothesis testing with the motion models to reassign the regions. We use a simple cost function, $C(i(x,y))$, that minimizes the velocity errors between the local motion estimates and the expected motion described by the affine models. This cost function is summarized as follows:

$$C(i(x,y)) = \sum_{x,y} (\mathbf{V}(x,y) - \mathbf{V}_{H_i}(x,y))^2 \tag{6}$$

where $i(x,y)$ is the indicates the model that location (x,y) is assigned to, $\mathbf{V}(x,y)$ is the estimated local motion field, and $\mathbf{V}_{H_i}(x,y)$ is the affine motion field corresponding to the i^{th} hypothesis. Since each location is assigned to only one of the hypotheses, we obtain the minimum total cost by minimizing the cost at each location. We summarize the assignment in the following equations:

$$i_0(x,y) \quad = \arg \quad \min[\mathbf{V}(x,y) - \mathbf{V}_{H_i}(x,y)]^2 \tag{7}$$

where $i_0(x,y)$ is the minimum costs assignment. Regions that are not easily described by any of the models are unassigned. These regions usually occur at object boundaries because the assumptions used by the optic flow estimation are violated. We assign these regions by warping the images according to the affine motion models and selecting the model that minimizes the error in intensity between the pair of images.

We now define the binary region masks that describe the support regions for each of the affine hypotheses as:

$$P_i(x,y) = \begin{cases} 1 & \text{if } i_0(x,y) = i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

These region masks allow us to identify the object regions and to refine our affine motion estimates in the subsequent iterations according to Equations 5.

As we perform more iterations, we obtain more accurate motion segmentation because the affine motion estimation is performed within single motion regions. Convergence is obtained when only a few points are reassigned or when the number of iterations reaches the maximum allowed. Models that have small support regions are eliminated because their affine parameters will be inaccurate in these small regions.

We maintain the temporal coherence and stability of the segmentation by using the current motion segmentation results as initial conditions for segmentation on the next pair of frames. Since an object's shape and motion change slowly from frame to frame, the segmentation results between consecutive frames are similar and require fewer iterations for convergence. When the motion segmentation on the entire sequence is completed, each object will have a region mask and an affine motion description for each frame of the sequence.

3.2 Analysis of layers

The images of the corresponding regions in the different frames differ only by an affine transformation. By applying these transformations to all the frames, we align the corresponding regions in the different frames. When the motion parameters are accurately estimated, objects will appear stationary in the motion compensated sequence. The layer images and opacity map are derived from these motion compensated sequences.

However, some of the images in the compensated sequence may not contain a complete image of the object because of occlusions. Additionally, an image may have small intensity variations due to different lighting conditions. In order to recover the complete representative image and boundary of the object, we collect the data available at each point in the layer and apply a median operation on the data. This operation can be easily seen as a temporal median filtering operation on the motion compensated sequence in regions defined by the region masks. Earlier studies have shown that motion compensation median filter can enhance noisy images and preserve edge information better than a temporal averaging filter [6].

Finally, we determine occlusion relationship. For each location of each layer, we tabulate the number of corresponding points used in the median filtering operation. These images are warped to their respective positions in the original sequence according to the estimated affine motions and the values are compared at each location. An layer that is derived from more points occludes an image that is derived from fewer points, since an occluded region necessarily has fewer corresponding points in the recovery stage. Thus the statistics from the motion segmentation and temporal median filtering provide the necessary description of the object motion, texture pattern, opacity, and occlusion relationship.

Our modular approach also allows us to easily incorporate other motion estimation and segmentation algorithm into a single robust framework.

4 Experimental results

We implemented the image analysis technique on a SUN workstation and use the first 30 frames of the MPEG "flower garden" sequence to illustrate the analysis, the representation, and synthesis. Three frames of the sequence, frames 0, 15 and 30, are shown in Figure 3. In this sequence, the tree, flower bed, and row of houses move towards the left but at different velocities. Regions of the flower bed closer to the camera move faster than the regions near the row of houses in the distance.

Optic flow obtained with a multi-scale coarse-to-fine gradient method on a pair of frames is shown on the left in Figure 4. The initial regions used for the segmentation consisted of 215 square regions. Notice the poor motion estimates along the occlusion boundaries of the tree as shown by the different lengths of the arrows and the arrows pointing upwards. In the same figure, results of the affine motion segmentation is shown on the middle. The affine motion regions are depicted by different gray levels and darkest regions along the edges of the tree in the middle figure correspond to regions where the local motion could not be accurately described by any of the affine models. Region assignment based on warping the images and minimizing intensity error reassigns these regions and is shown on the right.

Our analysis decomposed the image into 4 primary regions: tree, house, flower-bed and sky. Affine parameters and the support regions were obtained for the entire sequence, and the layer images for the four objects obtained by motion compensated temporal median filtering are shown in Figure 5. We use Frame 15 as the reference frame for the image alignment. The occluding tree has been removed and occluded regions recovered in the flower-bed layer and the house layer. The sky layer is not shown. Regions with no texture, such as the sky, cannot be readily assigned to a layer since they contain no motion information. We assign these regions to a single layer that describes stationary textureless objects.

We can recreate the entire image sequence from the layer images of Figure 5, along with the occlusion information, the affine parameters that describe the object motion, and the stationary layer. Figure 6 shows three synthesized images corresponding to the three images in Figure 3. The objects are placed in their respective positions and occlusion of background by the tree is correctly described by the layers. Figure 7 shows the corresponding frames synthesized without the tree layer. Uncovered regions are correctly recovered because our layered representation maintains a description of motion in these regions.

5 Conclusions

We employ a layered image motion representation that provides an accurate description of motion discontinuities and motion occlusion. Each occluding and occluded object is explicitly represented by a layer that describes the object's motion, texture pattern, shape, and opacity. In this representation, we describe motion discontinuities as discontinuities in object surface opacity rather than discontinuities in the actual object motion.

To achieve the layered description, we use a robust motion segmentation algorithm that produces stable image segmentation and accurate affine motion estimation over time. We deal with the many problems in motion segmentation by appropriately applying the image constraints at each step of our algorithm. We initially estimate the local motion within the image, then iteratively refine the estimates of object's shape and motion. A set of likely affine motion models exhibited by objects in the scene are calculated from the local motion data and used in a hypothesis testing framework to determine the coherent motion regions. Finally, the temporal coherence of object shape and texture pattern allows us to produce a description of the object image, boundary and occlusion relationship. Our approach provides useful tools in image understanding and object tracking, and has potentials as an efficient model for image sequence coding.

Acknowledgements

This research was supported in part by a contract with SECOM Co., and Goldstar Co., Ltd.

References

[1] E.H. Adelson, Layered representation for image coding, Technical Report No. 181, Vision and Modeling Group, The MIT Media Lab, December 1991.

[2] J. R. Bergen, P. J. Burt, R. Hingorani, and S. Peleg , Computing two motions from three frames, *International Conference on Computer Vision*, 1990.

[3] M. J. Black, P. Anandan, Robust dynamic motion estimation over time, *Proc. IEEE Computer Vision and Pattern Recognition91*, pp. 296-302, 1991.

[4] T. Darrell, and Alex Pentland, Robust estimation of multi-layered motion representation, *IEEE Workshop on Visual Motion*, pp. 173-178, Princeton, 1991.

[5] R. Depommier R., E. Dubois, Motion estimation with detection of occlusion areas, *Proc. IEEE ICASSP92*, Vol. 3, pp. 269-273, San Francisco, March 1992.

[6] T. S. Huang and Y. P. Hsu, "Image Sequence Enhancement," Image Sequence Analysis, Editor T. S. Huang, pp. 289-309., Springer-Verlag, 1981.

[7] M. Irani, S. Peleg, Image sequence enhancement using multiple motions analysis, *Proc. IEEE Computer Vision and Pattern Recognition92*, pp. 216-221, Champaign, June, 1992.

[8] Lucas, B., Kanade, T., An iterative image registration technique with an application to stereo vision, *Image Understanding Workshop*, pp. 121-130, April, 1981.

[9] S. Nagahdaripour, S. Lee, Motion recovery from image sequences using first-order optical flow information, *Proc. IEEE Workshop on Visual Motion 91*, pp. 132-139, Princeton, 1991.

[10] M. Shizawa and K. Mase, A unified computational theory for motion transparency and motion boundaries based on eigenenergy analysis," *Proc. IEEE Computer Vision and Pattern Recognition91*, pp. 296-302, 1991.

[11] C. W. Therrien, Decision Estimation and Classification, John Wiley and Sons, New York, 1989.

Figure 3: Frames 0, 15 and 30, of MPEG flower garden sequence.

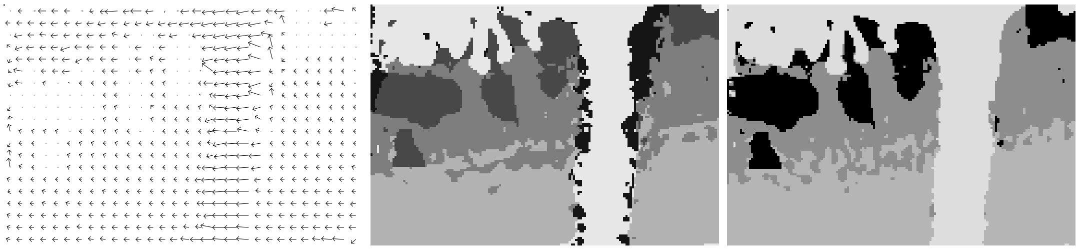

Figure 4: Affine motion segmentation of optic flow.

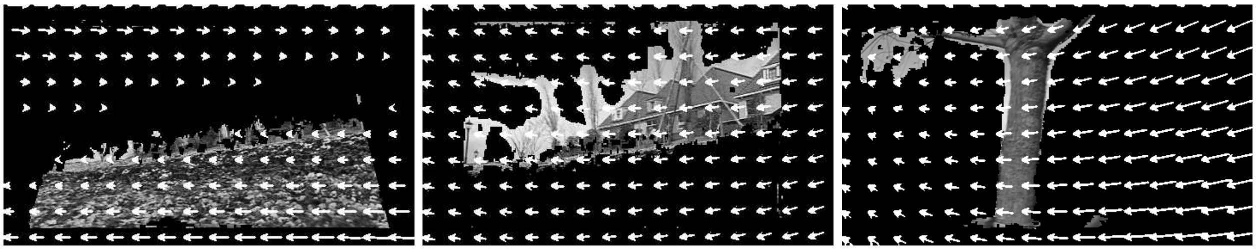

Figure 5: Images of the flower bed, houses, and tree. Affine motion fields are also shown here.

Figure 6: Corresponding frames of Figure 3 synthesized from layer images in Figure 5.

Figure 7: Corresponding frames of Figure 3 synthesized without the tree layer.