A New Bayesian Relaxation Framework for the Estimation and Segmentation of Multiple Motions

Alexander Strehl and J. K. Aggarwal *
Computer and Vision Research Center
The University of Texas at Austin
Department of Electrical and Computer Engineering
Austin, TX 78712-1084, U.S.A.
{strehl,aggarwaljk}@mail.utexas.edu

Abstract

In this paper we propose a new probabilistic relaxation framework to perform robust multiple motion estimation and segmentation from a sequence of images. Our approach uses displacement information obtained from tracked features or raw sparse optical flow to iteratively estimate multiple motion models. Each iteration consists of a segmentation and a motion parameter estimation step. The motion models are used to compute probability density functions for all displacement vectors. Based on the estimated probabilities a pixel-wise segmentation decision is made by a Bayesian classifier, which is optimal in respect to minimum error. The updated segmentation then relaxes the motion parameter estimates. These two steps are iterated until the error of the fitted models is minimized. The Bayesian formulation provides a unified probabilistic framework for various motion models and induces inherent robustness through its rejection mechanism. An implementation of the proposed framework using translational and affine motion models is presented. Its superior performance on real image sequences containing multiple and fragmented motions is demonstrated.

1 Overview

Visual motion is an important cue for a wide range of scene analysis tasks such as vehicle navigation, structure from motion, sequence stabilization, image segmentation or object tracking [10]. Visual motion may arise from moving objects viewed by a *fixed camera* against a static background, in which case the moving regions are used to identify and track objects [4]. Or, visual motion may imply a *moving camera* depicting a still background. [7]. The combination of a moving camera and moving objects is the *multiple motion* problem, in which segmentation and egomotion estimation have to be solved at the same time. Existing approaches for the analysis of *multiple motions* can be classified as those that use segmentation to separate motion [8] [11], and those that do not [3] [12]. In this paper we propose a new probabilistic relaxation framework to perform robust multiple motion estimation and segmentation from a sequence of images. Figure 1 gives a graphical overview of our proposed system. Our approach uses displacement information obtained from tracked features or raw sparse optical flow to iteratively estimate multiple motion models. Each iteration consists of a segmentation and a motion parameter estimation step. The motion models are used to compute probability density functions for all displacement vectors. Based on the estimated probabilities, a pixel-wise segmentation decision is made by a Bayesian classifier, which is optimal in respect to minimum error. The updated segmentation then relaxes the motion parameter estimates. These two steps are repeated until the error of the fitted models is minimized. Unlike region-based tracking methods, our system assumes neither coherently moving areas nor parametric shape approximations. This enables us to successfully analyze scenes containing multiple fragmented and occluded motions. Compared to popular segmentation-free approaches, which consider only the dominant motion, our framework simultaneously addresses multiple motions. The Bayesian formulation (section 2) provides a unified probabilistic framework for various motion models and induces inherent robustness against outliers through rejection of classifications with insufficient *a posteriori* confidence. An implementation of the proposed framework using translational and affine motion models is

*This research was supported in part by the Army Research Office under contracts DAAH04-95-I-0494 and DAAG55-98-1-0230 and by the Texas Higher Education Coordinating Board Advanced Research Project 97-ARP-275.

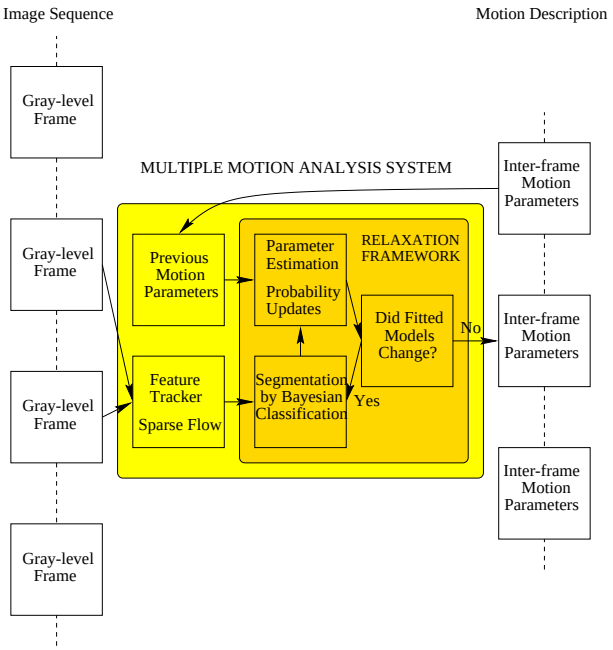

Figure 1. Overview of our proposed motion analysis system.

presented in section 3. Its superior performance on real image sequences containing multiple and fragmented motions is demonstrated. The estimated parameters are used to segment, stabilize and mosaic scenes showing the accuracy of the recovered motions (section 4).

2 Bayesian Relaxation Framework

Our proposed methodology works on displacement vectors that are computed by a windowed search for the maximum cross-correlation (MCC) [13] using a coarse-to-fine [2] framework. The framework segments this set of displacement vectors and estimates each segment's motion parameters. This is done in an iterative fashion, as illustrated in the dark box of figure 1. Each iteration consists of two steps:

- *Motion parameter estimation*
- *Segmentation*

First, the motion is estimated for each segment. This leads to an update of the likelihood for each location's displacement. A Bayesian classifier is employed to re-segment the image according to the new likelihood estimates. The new segmentation yields new and relaxed motion estimates. These two steps are iterated as long as the estimated parameters or the segmentation change. Upon convergence of this optimization, the current estimates have minimal error and constitute the final estimates. Since there is no proof of convergence at this time, the relaxation is *also* terminated when a certain number of iterations is exceeded. Subsection 2.1 discusses various methods to initialize the relaxation process and subsection 2.2 presents a detailed description of the two steps within each iteration. Section 3 gives details on how we implemented these two steps using an affine motion model.

2.1 Initialization

Like any iterative optimization paradigm, an initial guess is needed to start the optimization process. In our case, either an initial segmentation or an initial set of motion parameters is needed. The initialization is of crucial importance to avoid converging to a suboptimal (local) error minimum. For our framework, we considered the following three methodologies:

- *Random initialization* – All locations are assigned a random motion class label. We chose this to be our default initialization method for the first pair of frames if no *a priori* knowledge is given.
- *Initialization with previous results* – After processing the first image pair, previous motion parameters are used to provide a good initialization of the relaxation process of the current frame pair. This method infers a certain amount of motion smoothness over time. Due to mass inertia, this is a good assumption as long as inter-frame times are sufficiently small.
- *Domain specific initialization* – Environment information can also be utilized. In an active vision tracking setting, for example, we could assume that in the image plane, the object moves slower than the background, and thus initialize according to the velocity vector magnitude. Or we can exploit scene knowledge for initialization when, for example, the object is known to be found mostly in the image center.

2.2 Motion Parameter Estimation and Segmentation

Through the initialization, each location was assigned a preliminary (random) label segmenting the set of locations into multiple motions. Let s be the number of motions in the scene and t the number of parameters in each motion model. In the *motion estimation step*, the inter-frame parameter matrix θ (of size s by t) is computed from the set of displacement vectors. This task consists of independently estimating the parameter vector θ_i for each motion class m_i. Only locations labeled as a member of a certain motion m_i

contribute to the estimation of its parameter vector $\hat{\theta}_i$. How the displacement vectors of the form $(u, v)^T$ are used to estimate the parameter vector $\hat{\theta}_i$ depends on the selected motion model family. A generic 2–dimensional planar transformations model family is given by equations 4 and 5, and an implementation of maximum likelihood (ML) estimation is given in section 3. Closed form or iterative regression solutions for the motion model parameters $\hat{\theta}$ can be derived by following the ML approach for independent samples r_j as given by equation 1.

$$\hat{\theta}_i = \max_{\theta_i} \prod_j p(r_j|m_i, \theta) \tag{1}$$

The estimated motion parameters describe the fitted motion models. In the *segmentation step* each location r_j's motion class label l_j is updated. The new label l_j is determined by a Bayesian classifier [5]. The current motion models determine the likelihood of the image displacements. Hence, changing the motion estimates updates all class-specific probabilities $p(r_j|m_i, \theta)$ that the location r_j has a displacement vector as computed by the preprocessing stage, given that the correct motion type is m_i and the parameters θ are known. In order to implement the proposed framework, a parametric probability distribution family for $p(r_j|m_i, \theta)$ has to be selected (e.g., Gaussian, see section 3). Given $p(r_j|m_i, \theta)$ we can use Bayes' rule [5] to express the *a posteriori* probability $P(m_i|r_j, \theta)$ that a location r_j's displacement was generated by motion m_i, as follows:

$$\begin{aligned} P(m_i|r_j, \theta) &= \frac{p(r_j|m_i, \theta) \cdot P(m_i|\theta)}{p(r_j|\theta)} \\ &= \frac{p(r_j|m_i, \theta) \cdot P(m_i|\theta)}{\sum_{k=1}^{s} p(r_j|m_k, \theta) \cdot P(m_k|\theta)} \end{aligned} \tag{2}$$

Knowing the *a posteriori* probabilities, the optimal classification (in the sense of minimum misclassifications) is made with Bayes' decision rule [5], as follows:

$$l_j = \begin{cases} \text{argmax}_i(P(m_i|r_j, \theta)) & \text{if } \max_i(P(m_i|r_j, \theta)) > c \\ 0 & \text{else} \end{cases} \tag{3}$$

The label l_j can indicate a motion class $i > 0$ for $l_j = i$ or rejection if $l_j = 0$. The scalar variable $c \in [0, 1]$ denotes a user-selected threshold to control the amount of rejection. The maximum *a posteriori* (MAP) probability for the decision has to exceed the confidence *c*. For $c = 0$, there is no rejection, and for $c = 1$, every pixel is rejected. Useful values for c should be significantly greater than the blind guess probability $1/s$ and less than 1. (We used 0.9 for our experiments.) The higher the value of c is chosen, the more likely are rejections. The following section discusses an affine implementation of the two described abstract steps in our framework.

3 Affine Implementation

The planar model can describe the motion of a plane in 3–dimensional space projected into the 2–dimensional image [1] [9]. Let $\mathbf{M}_i^{a,b}$ denote a transformation that warps frame F_a to match frame F_b in respect to motion m_i. This transformation of location r_j's Cartesian coordinates $(x_j^a, y_j^a)^T$ can be conveniently written as a matrix multiplication using the corresponding homogeneous coordinates $\mathbf{f}_j^a$ in frame F_a to $\mathbf{f}_j^b$ in frame F_b [6]:

$$\mathbf{f}_j^b = \mathbf{M}_i^{a,b} \cdot \mathbf{f}_j^a \tag{4}$$

with $\mathbf{M}_i^{a,b}$ and $\mathbf{f}_j$ defined as

$$\mathbf{M}_i^{a,b} = \begin{pmatrix} \theta_{i,1}^{a,b} & \theta_{i,2}^{a,b} & \theta_{i,3}^{a,b} \\ \theta_{i,4}^{a,b} & \theta_{i,5}^{a,b} & \theta_{i,6}^{a,b} \\ \theta_{i,7}^{a,b} & \theta_{i,8}^{a,b} & 1 \end{pmatrix} \quad \mathbf{f}_j^a = w_j^a \begin{pmatrix} x_j^a \\ y_j^a \\ 1 \end{pmatrix} \tag{5}$$

In the following, we consider only one pair of images and consequently omit the temporal indices a and b for improved readability. Imposing the constraints $\theta_{i,7} = 0$ and $\theta_{i,8} = 0$ simplifies the planar motion model to the popular affine model. In this case, the displacements in the image are modeled as an affine transformation of their coordinates. Adding the restrictions $\theta_{i,1} = 1$, $\theta_{i,2} = 0$, $\theta_{i,4} = 0$ and $\theta_{i,5} = 1$ delivers the translational model that can only account for displacements constant throughout the image. The affine motion model showed the best results because it provides a good balance for the trade-off to keep model error and estimation error low.

Let us first look at the estimation step. Assuming that the affine motion model holds for all n_i locations r_j (that are labeled as motion m_i) gives us an over-determined system of n_i equations. A closed form least-squared-error (LSE) solution for $\hat{\theta}$ can easily be obtained using the pseudo-inverse method, and is given by

$$\begin{pmatrix} \hat{\theta}_{i,1} \\ \hat{\theta}_{i,2} \\ \hat{\theta}_{i,3} \end{pmatrix} = \begin{pmatrix} \sum x^2 & \sum xy & \sum x \\ \sum xy & \sum y^2 & \sum y \\ \sum x & \sum y & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum x^2 + xu \\ \sum xy + yu \\ \sum x + u \end{pmatrix} \tag{6}$$

and

$$\begin{pmatrix} \hat{\theta}_{i,4} \\ \hat{\theta}_{i,5} \\ \hat{\theta}_{i,6} \end{pmatrix} = \begin{pmatrix} \sum x^2 & \sum xy & \sum x \\ \sum xy & \sum y^2 & \sum y \\ \sum x & \sum y & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum xy + xv \\ \sum y^2 + yv \\ \sum y + v \end{pmatrix} \tag{7}$$

The summations are over all $j \in \{1, \ldots, n_i\}$ and the index j is omitted at all x, y, u and v symbols in equations 6 and 7. With these equations, the LSE estimates of the motion parameters can be computed from the displacement vectors $(u_j, v_j)^T$ and the preliminary classification. Simple LSE estimation is not robust, meaning that a single outlier

can arbitrarily worsen the estimate. However, our framework provides inherent robustness: Points that deviate too much from the current model estimate are labeled as outliers (rejection class 0) by the Bayes classifier and are excluded from the estimation procedure. The tolerance of a motion model m_i to deviations is specified in terms of the decision confidence threshold c and the tolerance matrix $\mathbf{C}_i$ as described below. In other words, the robustness lies in our framework and thus it is not necessary to implement robustness for every particular motion model estimator.

The second choice when implementing the framework is which error distribution model to use for the classification step. In other words, how do the pixel displacements estimated in the preprocessing stage deviate from the displacements predicted from a certain motion model m_i? We assumed the error of the displacement vectors to be signal-independent bivariate zero-mean Gaussian additive noise. The Gaussian model is a good first approximation and a well-known noise model. Using a Gaussian distribution, the class-specific probability $p(r_j|m_i,\theta)$ that a displacement at location r_j was generated by motion m_i is determined to be

$$p(r_j|m_i,\theta) = \frac{\sqrt{|\mathbf{C}_i^{-1}|}}{2\pi} \exp\left(-\frac{1}{2}\mathbf{d}_{i,j}^T \mathbf{C}_i^{-1} \mathbf{d}_{i,j}\right) \quad (8)$$

where location r_j's difference $\mathbf{d}_{i,j}$ between computed and predicted displacement is

$$\mathbf{d}_{i,j} = \begin{pmatrix} u_j \\ v_j \end{pmatrix} - \begin{pmatrix} \hat{u}_{i,j} \\ \hat{v}_{i,j} \end{pmatrix} \quad (9)$$

and location r_j's displacement $(\hat{u}_{i,j}, \hat{v}_{i,j})^T$ predicted from motion model m_i is given by

$$\begin{pmatrix} \hat{u}_{i,j} \\ \hat{v}_{i,j} \\ 0 \end{pmatrix} = (\mathbf{M}_i - \mathbf{I}) \cdot \begin{pmatrix} x_j \\ y_j \\ 1 \end{pmatrix} \quad (10)$$

The *a priori* probability $P(m_i|\theta)$ is the probability of encountering a displacement caused by motion m_i when picking a random location r_j from the image, given the parameters θ. Unless we have a specific and known scene model, the parameters of the considered 2–dimensional motion models are independent from the *a priori* probabilities, so $P(m_i|\theta) = P(m_i)$. In graphical terms, $P(m_i)$ represents the fraction of the image area that undergoes motion m_i. This parameter can be used to incorporate *a priori* knowledge about the analyzed scenes. For example, given the fact that in most cases the primary motion covers 80% of the image, $P(m_1)$ should be set to 0.8. We want to analyze scenes with two motions, namely object and camera motion, while not tuning our system to a certain object size. Consequently, given s motions, let $i \in \{1,\ldots,s\}$ and assume equal *a priori* probabilities $P(m_i) = 1/s$.

The covariance matrix $\mathbf{C}_i$ represents the tolerance of a motion class m_i regarding deviations of a single location r_j's computed displacement $(u_j, v_j)^T$ from the predicted displacement $(\hat{u}_{i,j}, \hat{v}_{i,j})^T$ according to a certain model m_i. This matrix can be used to encode the error characteristics of motion m_i. If the displacement estimates for a certain motion class are known to be affected by large errors, the covariance matrix can be set accordingly to tolerate the high deviations. Depending on the specific application of our framework, it may also be desirable to adjust $\mathbf{C}_i$ dynamically for each iteration step (making it another parameter or a function of the parameters) rather then using a fixed value for each class m_i. By default, we do not specialize on certain cases and chose equal tolerance for all motion classes. $\mathbf{C}_i$ is set to the 2 by 2 unity matrix $\mathbf{I}$ for all i. Now we have all the required information to start relabeling (classifying) each location r_j in the frame with the new label l_j, using Bayes' decision rule (equation 3). The two steps, segmentation and estimation, are iterated until the results stabilize, which typically happens after 2 to 20 iterations.

4 Experimental Results

Figures 2(a) and (b) depict the original two frames of the Skater sequence [3], an example of highly fragmented motion. It shows a crowd of people in the background (secondary motion) that is occluded by trees in the foreground (primary motion). Using our estimated parameters to align the images according to primary and secondary motion, the remaining residuals are shown in figures 2(c) and (d), respectively. For example, in figure 2(d) the images were registered in respect to the crowd's motion before differencing. Consequently the tree in the foreground is salient in the error image. Figures 2(e) and (f), respectively, show the successful motion-based segmentation into tree (primary) and crowd (secondary) as produced by the Bayesian classifier. Please refer to `http://strehl.com/research/` for more examples and detailed results on a variety of scenarios and numbers of motions.

5 Conclusions and Future Work

In this paper, we presented a new robust stochastic relaxation framework for the estimation and segmentation of multiple motions. Within this versatile framework, a Bayesian formulation is employed to solve the segmentation problem. Our approach allows the robust estimation of various models in multiple motion scenarios within a unified framework. We demonstrate an implementation of our framework to estimate and segment affine motions in gray-level image sequences. The implementation is tested on actual sequences of moving objects depicted by moving

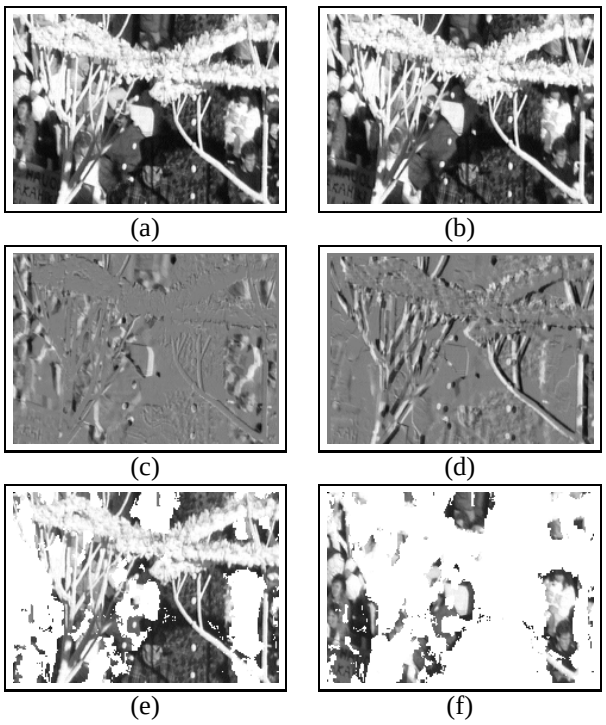

Figure 2. Fragmented motion in the Skater sequence.

cameras. Excellent results with low error are obtained for multiple motion and fragmented motion sequences.

In future work, we plan to experiment with a variety of parametric motion families (including 3–dimensional models). Another future direction could include experiments with simultaneous use of tracked feature points and sparse sub-pixel accurate optical flow. Moreover, an extension to the current framework could be developed to automatically pick the appropriate motion model for a scene based on an error measure obtained from the probability density functions of the Bayes classifier.

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(4):384–400, 1985.

[2] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model–based motion estimation. In *Proceedings European Conference on Computer Vision, Berlin, Germany*, volume 588 of *LNCS*, pages 237–252. Springer, May 1992.

[3] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, January 1996.

[4] Q. Cai, A. Mitiche, and J. K. Aggarwal. Tracking human motion in an indoor environment. In *Proceedings of the Second IEEE Conference on Image Processing*, pages 215–218, Washington D.C., 1995.

[5] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1972.

[6] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, Reading, MA, 1990.

[7] D. J. Heeger and A. D. Jepson. Subspace methods for recovering rigid motion: Algorithm and implementation. *International Journal of Computer Vision*, 7(2):95–117, 1992.

[8] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16, January 1994.

[9] M. Irani, B. Rousso, and S. Peleg. Recovery of ego-motion using image stabilization. In *International Conference on Computer Vision and Pattern Recognition*, pages 454–460, March 1994.

[10] A. Mitiche. *Computational Analysis of Visual Motion*. Plenum Press, 1994.

[11] C. H. Morimoto, D. Dementhon, L. S. Davis, R. Chellappa, and R. Nelson. Detection of independently moving objects in passive video. In *Proceedings of Intelligent Vehicles Workshop, Detroit, MI*, pages 270–275, September 1995.

[12] H. S. Sawhney, S. Ayer, and M. Gorkani. Mosaic based 2D&3D dominant motion estimation for mosaicing and video representation. In E. Grimson, editor, *International Conference on Computer Vision*, pages 583–590. IEEE, June 1995.

[13] Q. Wu. A correlation-relaxation-labeling framework for computing optical flow – template matching from a new perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(9):843–853, September 1995.