Adaptive Multi-Stage 2D Image Motion Field Estimation

Ulrich Neumann and Suya You

Computer Science Department
Integrated Media Systems Center
University of Southern California, CA 90089-0781

ABSTRACT

This paper addresses the problem of robust 2D image motion estimation in natural environments. We develop an adaptive tracking-region selection and optical-flow estimation technique. The strategy of adaptive region selection locates reliable tracking regions and makes their motion estimation more reliable and computationally efficient. The multi-stage estimation procedure makes it possible to discriminate between good and poor estimation areas, which maximizes the quality of the final motion estimation. Furthermore, the model fitting stage further reduces the estimation error and provides a more compact and flexible motion field representation that is better suited for high-level vision processing. We demonstrate the performance of our techniques on both synthetic and natural image sequences.

Keywords: motion estimation, optical flow, parametric model, augmented reality

1. INTRODUCTION

The estimation of inter-frame image motion is a key element in image sequence analysis. Almost all image sequence processing begins by attempting to find the motion vector field that describes how the images change with time. Accurate and robust image motion estimates are a prerequisite for a wide variety of applications ranging from dynamic scene analysis and interpretation, 3D motion and structure extraction to image stabilization and video data compression. In our work, we extend motion estimation techniques to the video-based augmented-reality-tracking problem [1]. By extracting a 2D-image motion field directly from a video stream, we estimate or constrain the real camera 3D motion parameters to maintain the dynamic alignment between a real and virtual camera, without depending on artificial landmarks (fiducials). To achieve this effect, the first critical step is to compute a robust 2D-image motion field that can be used to accurately recover the 3D motion parameters of the real camera. The accuracy of the motion field directly determines the accuracy of alignment between the real and virtual cameras.

Methods for computing 2D-image motion fields often rely on spatial and temporal gradients of the image intensity [2-8]. The estimation of a pixel motion is typical an ill-posed problem, and additional constraints are required. For example, a quadratic smoothness constraint or a higher-order spatial-temporal derivative constraint [1] can make the problem solvable. Most existing approaches operate on the entire image, intending to compute a motion vector for every pixel. It is difficult, however, except in trivial scene cases, to recover accurate and dense motion fields in a natural environment. Difficulties arise from scene geometry and texture complexity; variations in lighting and shading; multiple motions and occlusions. Unconstrained motion fields are typically not well described by any global model. Recent work on image motion estimation has focused on trying to find a balance between local dense motion estimates and global approaches[2]. The approach presented in this paper also concedes the goal of single global motion computation in exchange for local but reliable motion estimates. We

compute confidence measures to predict and verify the most reliable regions of estimated motion and model them with different parametric models.

In this paper, we present a new image motion estimation approach. Unlike conventional motion estimation methods that operate equally over the entire image space, our technique adaptively selects a set of image regions chosen for their suitability for good tracking and motion estimation. The adaptive selection of these regions is based on metrics that measure the confidence of tracking and motion estimation in each region. The benefits of this approach include reliable image motion estimation and efficient computation. The adaptive processes are data-driven from the input images, mimicking the natural eye movements to regions of interest in a scene. In order to increase the robustness and reliability of the motion estimation process, we propose a multi-stage estimation strategy. For every result, an evaluation is imposed to measure the confidence of the estimation. If the estimation error is large, the result may be iteratively updated to a better estimate. Regions with high confidence motion estimates are accepted while low confidence estimates cause regions to be discarded.

For many outdoor scene views, the distance between the camera and scene is relatively far so the 2D-image motion field can be modeled with a parametric model. Since the regions we use are relatively small, we can view the imaging procedure and image motion field as an affine procedure, and fit the motion field with an affine parametric model in each accepted region. This modeling process also contributes to a region's confidence measure.

The following sections detail our method. Section 2 reviews the differential-based local optical flow model we use for image motion estimation. It forms the basis of the adaptive region selection, and in Section 3 we describe the region selection metrics and procedures. Section 4 introduces the multi-stage motion-estimation strategy and shows how this approach improves motion estimates. Section 5 presents the modeling procedure for estimated motion fields, and finally, we show some experimental results and concluding remarks in Section 6.

2. DIFFERENTIAL-BASED LOCAL OPTICAL FLOW MODEL

The fundamental assumption in differential-based optical flow estimation is intensity conservation, from which the principal relationship between intensity derivatives and optical flow, the *optical flow gradient constraint equation* [3], can be derived

$$I(\mathbf{x},t)\ \ \mathbf{v} + I_t(\mathbf{x},t) = \mathbf{0} \tag{1}$$

Where $I_t(\mathbf{x},t)$ denotes the partial time derivative of $I(\mathbf{x},t)$, $I(\mathbf{x},t) = \left(I_x(\mathbf{x},t), I_y(\mathbf{x},t)\right)$, and $I\ \ \mathbf{v}$ denotes the usual dot product. Unconstrained general flow fields are typically not well described by any global model due to the complexity of the motion field. Different local models have been used to facilitate the estimation process, including constant flow with a local window and locally smooth or continuous flow. The former facilitates direct local estimation [8], whereas the latter model requires iterative relaxation techniques. In our work, we use the local constant model, because its results compare favorably with others [2]. In this model, the optical flow field is fit to a constant model in each small spatial neighborhood. Optical flow estimates are computed by minimizing the weighted least-squared fitting

$$E(\mathbf{x}) = \sum_{\mathbf{x}\ \Omega} W^2(\mathbf{x})\left[\ \ I(\mathbf{x},t)\ \ \mathbf{v} + I_t(\mathbf{x},t)\right]^2 \tag{2}$$

where $W(\mathbf{x})$ denotes a window function that gives more influence to data at the center of the neighborhood than at the periphery. Minimizing this fitting error with respect to $\mathbf{v}$ leads to the equation $E(\mathbf{v}) = \mathbf{0}$, from which the optical flow field may be obtained as

$$\mathbf{v} = [\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}]^{-1}\mathbf{A}\mathbf{W}^{\mathbf{2}}\mathbf{B} \tag{3}$$

where, for n points $\mathbf{x}_i \quad \Omega$

$$\mathbf{A} = [\quad I(\mathbf{x}_1, t), \cdots, \quad I(\mathbf{x}_n, t)]^{\mathbf{T}}$$

$$\mathbf{B} = -[I_t(\mathbf{x}_1, t), \cdots, I_t(\mathbf{x}_n, t)]^{\mathbf{T}}$$

and

$$\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A} = \sum_{\mathbf{x}\ \Omega} W^2(\mathbf{x}) \begin{bmatrix} I_x(\mathbf{x},t)^2 & I_x(\mathbf{x},t)I_y(\mathbf{x},t) \\ I_y(\mathbf{x},t)I_x(\mathbf{x},t) & I_y(\mathbf{x},t)^2 \end{bmatrix}$$

$$\mathbf{A}\mathbf{W}^2\mathbf{B} = \sum_{\mathbf{x}\ \Omega} W^2(\mathbf{x}) \begin{bmatrix} I_x(\mathbf{x},t)I_t(\mathbf{x},t) \\ I_y(\mathbf{x},t)I_t(\mathbf{x},t) \end{bmatrix}$$

The local model method repeatedly solves this 2×2 linear system and then integrates the computed values of $\mathbf{v}$ to provide an estimate of the optical flow field over the aggregate time interval.

3. Region Selection and Evaluation

Consider the above motion estimation equation (3). The system is closed form solvable only when the 2×2 matrix $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ is nonsingular. Since this matrix depends on the magnitudes of the spatial gradients and their range of orientations, we can reliably compute the value of optical flow at a point if the system represents suitable measurements at that point. Making some observations concerning the singularities of the system, we know that the rank of the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ will be full unless the directions of gradient vectors everywhere within the window coincide. It must be well conditioned, which means the eignvalues of $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ can not be too small and the differences between them must not differ by several orders of magnitude. We also note that, the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ is actually a covariance matrix of the image. It measures the correlation of image intensity, which can be a good indicator of the distributions of image structures over a small patch [9]. Summarizing the relationship between the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ and image structure, small eignvalues of $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ correspond to a relatively constant intensity within a region. A pair of large and small eignvalues corresponds to an uni-directional texture pattern, and two large eignvalues can represent corners, salt-and-pepper textures, or other patterns.

The eign-distributions of the covariance matrix $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ are used as a metric to select tracking regions in our system, and as a predictor of the confidence for the optical flow computation at a point. If two eignvalues of $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$ are greater than a pre-defined threshold, then the point is accepted as a good feature candidate.

$$\min(\lambda_{1,}, \lambda_2) > TH \tag{4}$$

After the feature candidates are selected by the above constraint (4), an evaluation value is assigned for each detected candidate to indicate its confidence of optical flow computing. Let λ be the minimum eignvalue of matrix $\mathbf{A}^{\mathbf{T}}\mathbf{W}^{\mathbf{2}}\mathbf{A}$, then the evaluation value assigned for each candidate is defined as

$$C = k\lambda \tag{5}$$

where k is a weighting coefficient. The detected feature candidates can be ranked according to their evaluation values, and a set of final good features $\{PF_i\}$ are selected as those that have the higher evaluation values (the number depends on various applications)

$$\{PF_i = \mathbf{x}_i(C) \mid i \in \textit{candidate set}, C > \textit{threshold}\} \tag{6}$$

The feature list is then used to select reliable regions for optical flow computing. A region is deemed reliable if it includes sufficient good features. In our implementation, the image is first divided into non-overlapping regions R_i. The number of good points per candidate region is tabulated and we select the N "best" regions $\{RF_i\}$ as

$$\{RF_i = R_i(S) \mid i \text{ ë } \textit{candidate region}, S > \textit{threshold}\} \tag{7}$$

where the metric of quality is *S*, given by

$$S = \frac{N_g}{N_T} \tag{8}$$

and N_g is the number of good points within the region, and N_T is the number of pixels in the region. The number of selected regions is arbitrary, depending on the complexity of the scene structure and the application, but larger numbers will result in greater computation time.

4. Multi-Stage Optical Flow Estimation

Any imaging distortion, especially in the natural environment, may have significant effect on the feature appearance that results in unreliable optical flow. To achieve robust motion estimation, we use a multi-stage computing strategy. The main idea of this strategy is to utilize the current estimate to create a confidence frame. The confidence frame is dynamically fed back to the estimation function to verify and evaluate the current estimate. The confidence frame models the motion deformation parameters, so it can effectively compensate for those effects.

The three basic components of the multi-stage strategy are image warping, motion residual estimation, and motion refinement. Let $R(\mathbf{x},t)$ be a region that is selected in the frame t. $F(\mathbf{x},t)$ is its optical flow field estimated for each feature point within the region. Based on the estimated optical flow field, a new region $R_c(\mathbf{x},t)$ can been reconstructed by warping the region $R(\mathbf{x},t)$ towards its target $R_t(\mathbf{x},t)$. This newly constructed region is called a confidence frame. Since the motion field is sparse, bilinear interpolation is applied to un-estimated points. Although interpolation may introduce errors in reconstruction, it produces a dense and reasonably accurate optical flow field in many cases. Our experiments indicate that the interpolated optical flow fields based on sparse but reliable estimates are more accurate than unreliable dense estimates.

Since the confidence frame $R_c(\mathbf{x},t)$ is generated directly from its estimated motion, it can be used as a measurement of the optical flow accuracy within the region. A motion residual is computed based on the least-squares distance between the confidence frame $R_c(\mathbf{x},t)$ and its target $R_t(\mathbf{x},t)$,

$$\varepsilon = \frac{\left\| R_t(\mathbf{x},t) - R_c(\mathbf{x},t) \right\|^2}{\max\{\left\| R_t(\mathbf{x},t) \right\|^2, \left\| R_c(\mathbf{x},t) \right\|^2\}} \tag{10}$$

The residual measures the error of the optical flow estimate in the current region of the current frame. If the error is large, the current region under-went significant motions or distortions that are not captured in the current motion estimate. The estimate may be refined by re-computing the motion between $R_c(\mathbf{x},t)$ and $R_t(\mathbf{x},t)$. The process is repeated until the estimation error converges to within a pre-defined threshold. If the motion residual remains above-threshold after a fixed number of iterations, the region is eliminated from the feature list, and a new selection process can be started.

5. Optical Flow Field Fitting

For many outdoor scene views, the distance between the camera and scene elements is relatively large so the 2D-image motion field can be modeled with a parametric model. Since the regions we use are relatively small, an

affine parametric model can well approximate the image motion of the region. The six-parameter affine model is defined as

$$\begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \begin{bmatrix} v_2(x - x_c) + v_3(y - y_c) + v_1 \\ v_4(x - x_c) + v_5(y - y_c) + v_6 \end{bmatrix} \tag{11}$$

where $\mathbf{u}(\mathbf{x}, \mathbf{a}) = [u(x,y), v(x,y)]^{\mathbf{T}}$ are the horizontal and vertical components of the image velocity at the image point $\mathbf{x} = [x, y]^{\mathbf{T}}$, and $\mathbf{v} = [v_1, v_2, v_3, v_4, v_5, v_6]^{\mathbf{T}}$ denotes the vector of parameters to be estimated relative to a region center (x_c, y_c).

This affine model is capable of representing the 2D projection of the 3D transformation and deformation of a planar surface, and approximates the same 3D motion of a distant object having a small 2D projection in the image. Each region's estimated optical flow field is fit to the model. If the fitting error is large, our model does not approximate the local motion field in this region, so we exclude the corresponding region from further tracking. If the error is small, we accept the estimation as the correct result, and the six model parameters are used to represent the motion field of the entire region.

6. Experimental Results and Conclusions

The approach described above has been implemented and integrated in our linear motion estimation system for augmented reality applications [11]. Here we show the results of applying the approach to some synthetic image sequences and live video sequences. To evaluate qualitative performance, we use the angle error measure [2, 7] and standard RMS error measure. The angle error measure treats image velocity as a spatio-temporal vector $\mathbf{v} = (u, v, 1)$ in units of (pixel, pixel, frame). The angular error between the correct velocity $\mathbf{v}_c$ and the estimate $\mathbf{v}_e$ is defined as

$$Error_{angle} = \arccos(\mathbf{v}_c \quad \mathbf{v}_e) \tag{12}$$

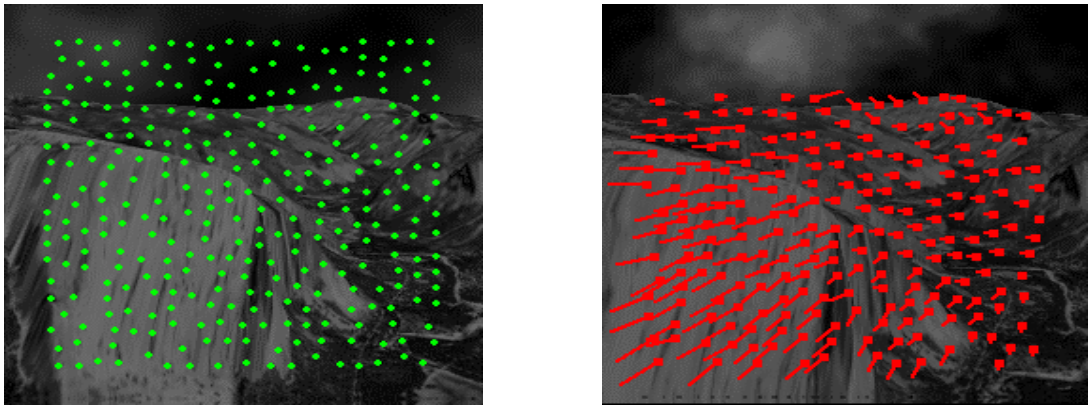
where $\mathbf{v}_i = \frac{(u, v, 1)^T}{\sqrt{u + v + 1}}$. This angle error measure is convenient, because it can handles large and small speeds without the amplifications inherent in a relative measure of vector difference. The measure also has potential bias, for example, directional errors at small velocity do not give as large an angular error as similar directional error at large velocity. For these reasons, we also used the RMS measure

$$Error_{rms} = \sqrt{\frac{\sum_{\mathbf{x} \in \Omega} (I_c(\mathbf{x}, t) - I_e(\mathbf{x}, t))^2}{MN}} \tag{24}$$

where $I_c(\mathbf{x}, t)$ is a size $M \times N$ region of a real image sequence at time t, and $I_e(\mathbf{x}, t)$ is the reconstructed region based on the estimated motion field. Note that this error measure is similar to the motion residual measure we used in section 4.

Figure 1 illustrates an experimental result applying our approach to the Yosemite-Fly-Through image sequence. The sequence is graphically generated by moving a camera along its line of sight towards the mountain and valley, generating diverging motion flow around the upper right of mountain. A 1 pixel/frame translating motion is obtained in the cloud area, and about 4 pixels/frame are found in the lower left portion. In this test, only one image region is selected whose size is the same as the original image size (256x256). In the region, 50% of the pixels with top ranked evaluation values are selected as features. We chose this test to facilitate performance comparisons with other optical flow approaches that compute motion estimates on whole images. Figure 1 (a) shows the selected feature points, and Figure 1 (b) illustrates the estimated optical flow field at the 15th frame. The resulting average angle error is 3.44 degrees, and the RMS error measure is 7.31. We compared our results

with those reported for some other optical flow approaches on the same test sequence [2-8]. The evaluation results for the Yosemite-Fly-Through sequence are given in Table 1.



(a) (b)

Figure 1. Yosemite-Fly-Through sequence: (a) detected tracking features, and (b) estimated motion field.

Technique	Average Angle Error	Standard Deviation
Horn and Schunck	11.26	16.41
Lucas and Kanade	4.10	9.58
Anandan	15.84	13.46
Fleet and Jepson	4.29	11.24
Proposed approach	3.44	10.28

Table 1: Performance comparison of various optical flow approaches

A second experiment uses a real natural image sequence, and an example image is shown in Figure 2. This sequence includes four objects moving against a static street background, with different motion directions and velocities. In this test, three regions are selected for motion estimation. The sizes of these regions are 61x61, and in each region, about 10 top-ranked points are selected for motion tracking (Figure 2 (a)). It is interesting to note that the distributions of the detected features are basically located around the significant physical features in the scene, such as object corners and edges, although we do not manually locate them. Normally these types of physical features are expected to be reliable for tracking, as noted in many publications, but our approach selects them based on the tracking metrics, as well as their spatial characteristics. Table 2 gives the RMS estimate error after applying our tracking approach.

(a) (b)

Figure 2. Hamburg Taxi Sequence: (a) first frame with detected region and point features, and (b) motion results at 20th frame.

Sequences	RMS
Yosemite-Fly-Through	7.31
Hamburg Taxi	4.54
Live video	11.03

Table 2: RMS errors for different image sequences by applying the proposed approach.

Finally, figure 3 shows the results of a video sequence captured with a CCD camcorder in a moving vehicle while viewing to the right and panning the camera. The computed RMS error result is also given in Table 2.

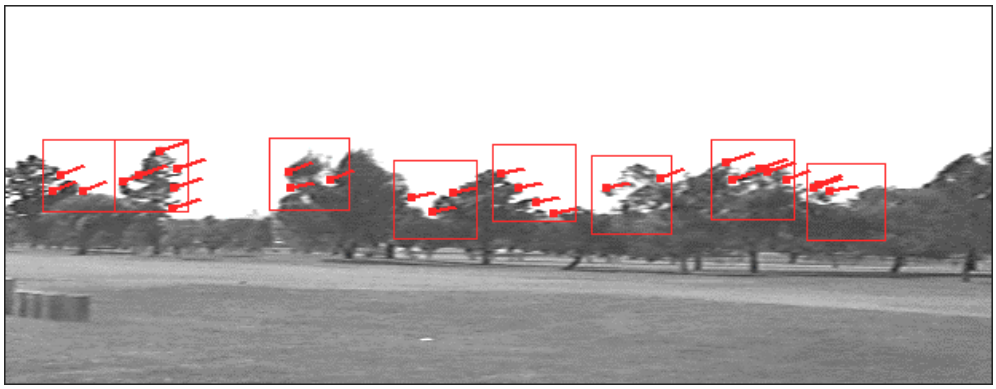


Figure 3. Result of an outdoor natural scene

In summary, to increase the robustness and efficiency of motion estimation in natural environments, we developed an adaptive region selection and optical flow estimation method. The strategy of adaptive region selection locates reliable tracking regions and makes their motion estimation more reliable and computationally efficient. This approach appears to have advantages over conventional methods. The multi-stage estimation procedure makes it possible to discriminate between good and poor estimation areas, and allows the system to maximize the quality of the final motion estimation. Furthermore, the model fitting stage reduces the estimation error for the entire region and provides a more compact and flexible motion field representation that is better suited for high-level vision processing.

Acknowledgments

This work has been supported by the Defense Advanced Research Project Agency (DARPA) "Geospatial Registration of Information for Dismounted Soldiers."

References

1. U. Neumann and J. Park. Extendible Object-Centric Tracking for Augmented Reality. *Proc. of IEEE Virtual Reality Annual International Symposium*, 148-155, 1998.

2. S. S. Beauchemin and J. L. Barron. The Computation of Optical flow. *ACM computing surveys*, 27(3):433-466, 1995.

3. B. K. P. Horn and B. G. schunk. Determining Optical Flow. *Artificial Intelligence*, 17:185-203.

4. H. H. Nagel. On a Constraint Equation for the Estimation of Displacement Rates in Image Sequences. *IEEE Trans. PAME*, 11:13-30, 1989.

5. E. C. Hildreth. Computation Underlying the Measurement of Visual Motion. *Artificial Intelligence*, 23:309-354, 1984.

6. P. Anandan. A Computational Framework and an Algorithm for the Measurement of Visual Motion. *Int. J. Comp. Vision*, 2:283-310, 1989.

7. D. J. Fleet and A. D. Jeson. Computation of Component Image Velocity from Local Phase Information. *Int. J. of Computer Vision*, 5:77-104, 1990.

8. B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. *Proc. DARPA IU Workshop*, 121-130, 1981.

9. C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method. Technical Report, Carnegie Mellon University, Pittsburgh, PA, September 1990.

10. G. D. Hager and P.N. Belhumeur. Real-Time Tracking of Image Regions with Changes in Geometry and Illumination. *IEEE Proc. of CVPR*, 1996.

11. U. Neumann and S. You. Integration of Region Tracking and Optical Flow for Image Motion Estimation. *International Conference on Image Processing (ICIP)*, 1998.