Non-Iterative Approach to Multiple 2D Motion Estimation

Eun-Young Kang, Isaac Cohen, Gerard Medioni
University of Southern California
Los Angeles, California 90089-0273
{elkang|icohen|medioni}@usc.edu

Abstract

We present an innovative method estimating multiple 2D motions from uncalibrated images. Our approach robustly and non-iteratively estimates multiple 2D parametric motions, affine or homography, from noisy initial matches without pre-specifying the number of motions This approach is based on: (1) a parametric motion model to detect and extract 2D affine or homography motions; (2) the representation of matching points in decoupled joint image spaces; (3) the characterization of the property associated with affine transformation in the defined spaces; (4) a non-iterative process to extract multiple 2D motions simultaneously based on tensor-voting; (5) local affine to global homography estimation. The major contribution of our work is the extension to our existing affine motion estimation method for homography estimation. The robustness of the approach is demonstrated with several results.

1. Introduction

Motion estimation along with grouping is essential for video processing. In particular, motion estimation using 2D parametric models is widely used for many applications such as video compression and video surveillance. In this paper, we present a robust and non-iterative method to estimate multiple 2D parametric motion groups. The major computational issues in order to extract multiple motion groups are (1) the determination of the number of motion groups, (2) the detection and removal of noisy mismatches, and (3) the clustering of matches. Here, we estimate multiple 2D motion groups without pre-specifying the number of motions in a robust and non-iterative manner. Our approach is based on following: 1) a parametric model to detect and extract 2D affine or homography motions, 2) the representation of the matching points in decoupled joint image spaces, 3) the characterization of the property associated with affine transformation in the defined spaces, 4) a process to extract multiple affine motions simultaneously based on tensor-voting, and 5) local affine to global homography estimation.

Many researchers have showed strong interests in multiple motion estimation and grouping over a decade and they have demonstrated many approaches. However, most of the methods use a pre-specified number of motion groups, and are sensitive to initial setting or noise. Parametric-based approaches extract motion groups by using a specific motion model to define the grouping constraint. In [5][12][14] the proposed methods are iterative. These iterative approaches either start with a pre-specified number of motions to fit pixels motions, or perform repetitive extraction-and-removal of dominant motions. In other words, they are based on an assumption of a known number of motions or an existence of conspicuously dominant motion. Regularization-based methods and Markov Random Fields (MRF) based methods focus on the discontinuity around motion boundaries or optical flow [3][6]. These approaches give some good results but they detect motions based on unreliable discontinuity around motion boundaries or rely on a spatial segmentation of the image, which may not be practical in many cases. To address the problems associated with multiple motions, several EM-based methods were proposed [1][2][4][11]. EM-based methods have showed good results. However, these EM-based methods also require specifying the number of regions and a significant amount of iteration to reach satisfactory performance. When EM-based methods are used without a specified number of motion groups, they usually require massive search, which might not guarantee the optimal grouping of regions. Tensor Voting–based method provides a non-iterative alternative by grouping regions using smoothness constraints in various space representations. For example, the estimation of the fundamental matrix parameters between two images [10]. These methods use tensor voting as a pre-processing step for outlier removal in 8D and 4D, which requires other iterative processing steps to determine different motion groups. As opposed to parametric approach, non-parametric and non-iterative methods were proposed [9][13]. In general, non-parametric approaches aim to provide common motion clustering. However, the applications requiring parameter recovery, such as compression, benefit from a motion segmentation approach that groups pixels into regions of similar parametric motions.

2. Detection of multiple affine motions

The work described in this paper extends the work presented in [7]. The method in [7] estimates multiple affine motions simultaneously by characterizing properties of motion groups, identifying inliers (correct matches) and outliers (mismatches), and assigning inliers into a proper

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

motion groups based on inferred properties. The method uses a decoupled joint image space to identify affine motion using tensor voting and estimates multiple 2D motions simultaneously. The tensor voting method allows extracting locally salient motion groups while removing outliers.

2.1. Affine motion in a decoupled image space

2D affine motion is defined by six parameters, the corresponding affine joint image space is defined in the 4D space *(x, y, x',y')* where *(x,y)* and *(x',y')* are an image correspondence. In the affine model, the two equations are independent, and therefore the joint space *(x, y, x',y')* can be decoupled into two joint spaces *(x,y,x')* and *(x,y,y')*. By defining $p_x = (a,b,-1,t_x)^T$ and $p_y = (c,d,-1,t_y)^T$, we have two separate joint spaces $q_x = (x,y,x')^T$ and $q_y = (x,y,y')^T$. We obtain the following equations in the decoupled joint image spaces:

$$p_x^T \begin{pmatrix} q_x \\ 1 \end{pmatrix} = ax + by - x' + t_x = 0,\ p_y^T \begin{pmatrix} q_y \\ 1 \end{pmatrix} = cx + dy - y' + t_y = 0$$

These equations define two plane equations in decoupled joint image spaces. A decoupled joint image space is a 3D space and a 2D plane in this space represents an affine motion. Therefore, by detecting 2D planes from input correspondences, we identify affine motion groups, extract groups of inliers, and estimate affine parameters directly from inliers. In reality, the sparse/dense pixel correspondence contains many mismatches. Therefore, we need to define a robust method for inferring salient plane features within the noisy input correspondences. Our approach uses tensor voting for achieving such selection.

2.2. Tensor voting for affine motion

Tensor voting formalism provides a robust approach for extracting salient structures, such as curve and surface [8]. The input data is encoded as second order symmetric tensors to capture first order differential geometry information and its singularities, and then each input (called *token*) propagates its information to the neighboring data by voting. Voting for neighbourhood is performed through voting fields that are generated based on a scale factor σ (size of neighbourhood). Vote orientation corresponds to the smooth local curve continuation from a voter to a recipient. The extraction of salient structures is inferred from the canonical description of an arbitrary tensor by its eigen-system representing the local geometric properties.

The considered input are point correspondences between two images. These correspondences are generated by cross correlation estimation. Then, each correspondence is normalized and is encoded into generic tensor form in 3D. In a 3-D case, a tensor is represented by an ellipsoid. In a more compact form, $\boldsymbol{S} = \lambda_1 \hat{\boldsymbol{e}}_1 \hat{\boldsymbol{e}}_1^T + \lambda_2 \hat{\boldsymbol{e}}_2 \hat{\boldsymbol{e}}_2^T + \lambda_3 \hat{\boldsymbol{e}}_3 \hat{\boldsymbol{e}}_3^T$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ are the eigenvalues, and $\hat{\boldsymbol{e}}_1$, $\hat{\boldsymbol{e}}_2$, $\hat{\boldsymbol{e}}_3$ are the eigenvectors corresponding to λ_1, λ_2, λ_3 respectively. The eigenvectors represent the principal directions of the ellipsoid and the eigenvalues encode the size and shape of the ellipsoid. We encode initial correspondences as an ellipsoid with equal eigenvalues for each eigenvector that propagates point information to all direction in any dimension. Once we have all initial tensors, we perform a voting for finding inliers.

Tensor voting is performed in two stages. First stage is used to infer preferred direction and then stronger support is collected in the second stage. The saliency of a point is characterized based on the plane normal saliency. Saliency is used to determine inliers and outliers. Here, we update voting process for better plane detection. First, we use different voting fields than usual. Original voting fields use smooth local curve continuation from a voter to a recipient. In our work, the field propagates the planar continuation from a voter to a recipient because we are looking for salient planes. This plane-fitting field enforces the affine metric during voting stage. Secondly, our approach utilizes boundary information during the first voting stage. The contour information is considered to inhibit voting between different colors regions or across boundaries and therefore it encourages voting within the same color regions.

2.3. Multiple motion detection

The motion grouping is performed by clustering points from the most salient points inferred by voting. This is a seed-based region growing method. However, our approach does not require specifying the number of motions in advance, and the seeds are not selected randomly. The seeds are determined by the inferred saliency from voting process, and their saliencies retain the likelihood measures. The coherence is based on plane-smoothness in the decoupled joint image space and is measured by the angular difference of the normal directions (characterized by the affine coefficients *a, b, c, d*) and the distance from the space's origin (characterized by the translation parameters *Tx* and *Ty*). At each point, the normal orientation of the plane is encoded by the eigenvector e_1 associated to the largest eigenvalue. We compare and cluster neighboring points according to their likelihood. This likelihood measure is provided by the distance functions: $\|d_s - d_i\|^2$, where $d_s = e_{11}^s x_s + e_{12}^s y_s + e_{13}^s x_s'$, $d_i = e_{11}^i x_i + e_{12}^i y_i + e_{13}^i x_i'$, *($x_s$, y_s, x_s')* and *(x_i, y_i, x_i')* are locations of the seed point, p_s and neighbor point, p_i, and each vector *(e_{11}, e_{12}, e_{13})* represents the normal direction of the plane at that location. This function approximates Euclidean distance between two parameterized motions. If motion difference is smaller than 1 pixel, two points are clustered together. After one region is clustered, the clustering procedure iterates using the next highest salient points. For each clustered motion group, the affine parameters are estimated by analyzing the correlation matrix of the clustered points in each space. The parameters of the affine transform are therefore characterized by the eigenvector associated to the smallest eigenvalue of the correlation matrix.

We have tested our approach on synthetic and real images. The results in Figure 1 and Figure 2 show accurate performance of our approach. In Figure 1, the synthetic motions are generated with 320x240 image size, three synthetic affine motions, total 874 correspondences, and 100% noise ratio. The result shows that measured average error between estimated affine parameters and synthetically generated parameters for three motion groups is less than 0.1 pixels. In Figure 2, we applied our approach in order to detect dominant motions from a video sequence and generate a panoramic view. The results in Figure 2 (b) show that our approach estimates more accurately the parameters than an iterative method based on RANSAC does.

3. Affine to homography

When scenes contain plane motions that are not parallel to the image plane and that have large depth variations, our affine motion extraction method detects several local motion patches for a motion of one plane. This is due to significant motion differences that cannot be modeled by a single affine motion. For these scenes, affine motion is not sufficient to approximate appearing motions. Therefore, we extend the local affine motion estimation method to 2D homography motion estimation. Images of points on a plane are related to corresponding image points in a second view by a planar homography. In order words, corresponding pixels in two images, which are projections of a world point on a plane in 3D space, are determined by a plane homography:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{31} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, x' = \frac{x_1}{x_3}, y' = \frac{x_2}{x_3}$$

We group affine patches into the same homography by using plane induced homographies given fundamental matrix F and image correspondences. When two images are rectified using F, each world plane forms a plane in 2.5D disparity space. Also, a motion of a world plane appearing two images is constrained by a homography. Therefore, by finding unique planes in 2.5D space, we can estimate unique homography groups. The first step is to find planes in 2.5D space, fundamental matrix F is computed from locally extracted motion groups. F is computed by least square method. We emphasize here that the matches used for estimating fundamental matrix are locally salient inliers only, and therefore F matrix estimation step is more stable. As the next step, we compute rectification homography based on F. A fixed number of epipolar lines are estimated from the fundamental matrix. These epipolar lines are used to rectify the images. After computing the rectification homography, we compute a normal direction for each affine patch. The computation of normal directions is based on the extracted affine parameters and rectification homography parameters. Here, we do not re-compute matches. When a space is defined by (*x, y,f(x,y)*), the direction of normal for a point in this space is defined by $N = (\partial f(x,y)/\partial x, \partial f(x,y)/\partial y, -1)^T$. When images are rectified with respect to the camera motion (expressed in F matrix), the normal direction of rectified affine patches in the disparity space leads to a homography group. The computation of the normal direction of a plane in 2.5D is dependent on the *f(x,y)*, which defines disparity function. In our case, it is derived from extracted affine parameters and rectification homography parameters. After each normal direction for each patch is computed, the grouping process is performed. If each estimated normal direction is similar (threshold value is 10^o), two patches are grouped into a single homography, which corresponds to a world scene plane or planes forming the same angle with the image plane.

4. Experimental Results

Let us consider images containing three planar objects (3 books). The images in Figure 3 (a) are taken by a hand-held camera (moved from left to right). In scenes, there exist several homographies corresponding to several planar objects (books). Our approach extracted 33 local affine groups. Several affine groups belong to a book (see Figure 3). Due to the difference of the motion magnitudes. In Figure 3 (c), three affine groups among those extracted affine groups are shown. Based on matches from extracted affine groups, fundamental matrix F and rectification homographies are computed. Then, for each affine patch, the new normal direction of the patches is computed by using the rectification homography and the affine motion parameters. Figure 3 (b) shows the plotted normal directions in the disparity space. Clearly, only three distinctive normal directions can be observed, and these three normals correspond to three world planes/books in the original input images. We can observe that the 33 affine groups belong to one of these groups. Figure 3 (d) shows grouping results. Given 10 degree threshold value, 3 homography groups are successfully detected. In Figure 3 (e), residual images are presented based on the extracted homography parameters. The original image is warped and compensated according to each estimated homography motion. Notice in Figure 3 (d) that our method successfully detected that the side of 'office-depot' box has same motion as the homography group of 'math' in the left.

5. Conclusion

This paper presented an innovative method to extract multiple 2D parametric motion groups, affine or homography, from noisy initial matches in *robust* and *non-iterative* manner. The major contribution of our paper is that we extended the existing affine motion estimation method to homography motions. We demonstrated that our approach gives promising results for estimating multiple object motions, which can be modeled by 2D transformations locally. In the future, further evaluation is

planned to be performed. A quantitative comparison to previous approaches, such as EM-based, will be addressed in terms of accuracy and performance.

Acknowledgements

This research was partially funded by the Advanced Research and Development Activity of the U.S. Government under contract MDA-904-03-C1786.

6. References

[1] S. Baker et al., A Layered Approach to Stereo Reconstruction, CVPR, 1998.

[2] M. Ben-Ezra, S. Peleg and M. Werman, Efficient Computation of the Most Probable Motion from Fuzzy Correspondences, IEEE Workshop on Applications of Computer Vision (WACV), 1998.

[3] Y. Boykov, O. Veksler, R. Zabih, Markov Random Fields with Efficient Aproximations, CVPR, pp. 648-655, 1998.

[4] T. Darrell and A. Pentland, Cooperative Robust Estimation Using Layers of Support, M.I.T Media Vision and Modeling Group Tech Report, No 163, Feb., 1991.

[5] M. Irani and S. Peleg, Motion Analysis for Image Enhancement: Resolution, Occlusion, and Transparency, Journal of Visual Communication and Image Representation (VCIR), Vol. 4, No. 4, pp. 324-335, December 1993.

[6] S. X. Ju, M. J. Black and A. Jepson, Skin and Bones: Multi-layer, Locally Affine, Optical Flow and Regularization with Transparency, Proceedings of CVPR, 1996.

[7] E.Y. Kang, I. Cohen and G. Medioni, Robust Affine Motion Estimation in Joint Image Space using Tensor Voting, International Conference on Pattern Recognition (ICPR), 2002.

[8] G. Medioni, et al., Tang, A Computational Framework for Feature Extraction and Segmentation, Elsevier, 2000.

[9] M. Nicolescu and G. Medioni, Perceptual Grouping from Motion Cues Using Tensor Voting in 4-D, ECCV, 2000, vol. III, Copenhagen, Denmark, May 2002, pp. 423-437.

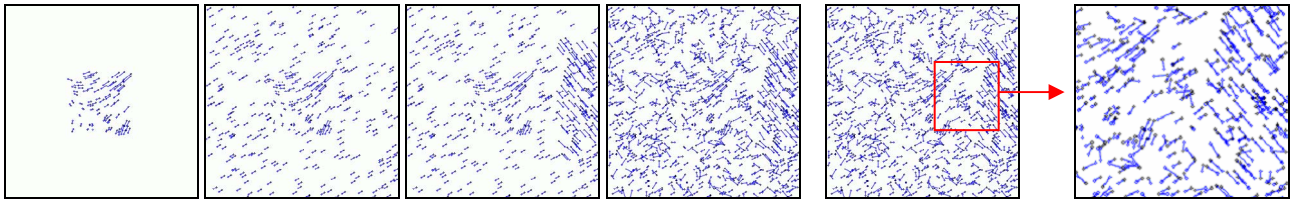
[10]] W. Tong, C. Tang and G. Medioni, Epipolar Geometry Estimation for Non-Static Scenes by 4D Tensor Voting, IEEE International Conference on Computer Vision (ICCV), 2001.

[11] A. P. Torr, R. Szeliski and P. Anandan, An Integrated Bayesian Approach to Layer Extraction from Image Sequences, IEEE Trans. on PAMI, 23(3), March 2001, pp. 297-303.

[12] J. A Wang and E. Adelson, Representing Moving Images with Layers, IEEE Transaction on Image Processing Special Issue: Image Sequence Compression, 3(5), September 1994.

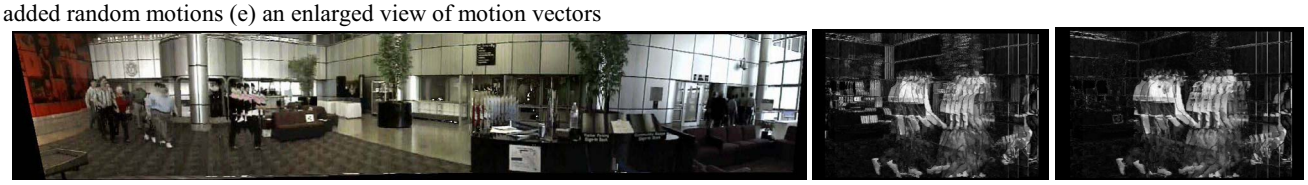
[13] Y. Weiss, Smoothness in Layers: Motion segmentation using nonparametric mixture estimation, ", CVPR, 1997,.

[14] L. Zelnik-Manor and M. Irani, Multi-Frame Estimation of Planar Motion, IEEE Trans. on PAMI, Vol. 22, No. 10, October 2000.



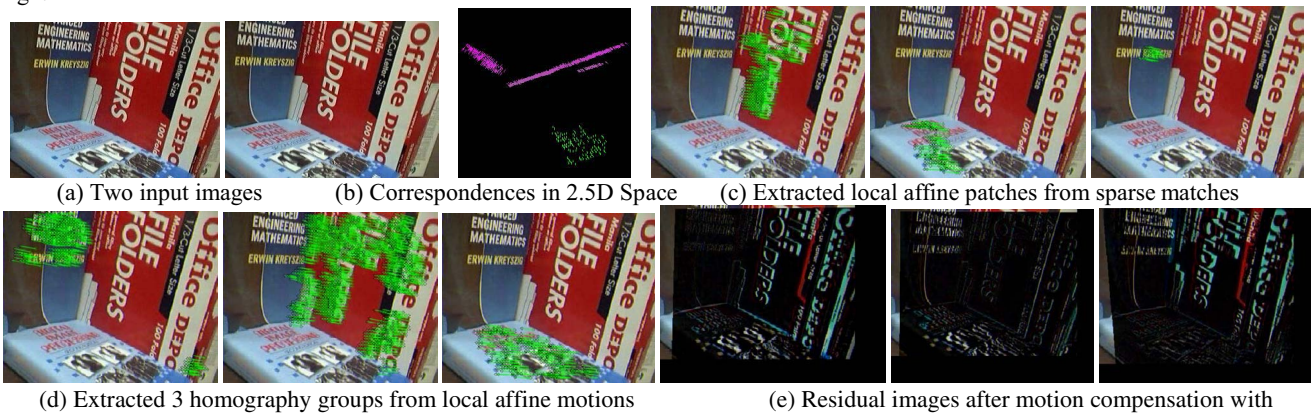
(a) (b) (c) (d) (e)

Figure 1: Synthetic motions with 100% noise (a) an affine motion (b) an added translation (c) an added transparent affine motion (d) added random motions (e) an enlarged view of motion vectors



(a) Panoramic image (b) residuals

Figure 2: Application to image mosaic. Residual by RANSAC-based iterative approach in (b) left and residual by our approach (b) right



(a) Two input images (b) Correspondences in 2.5D Space (c) Extracted local affine patches from sparse matches

(d) Extracted 3 homography groups from local affine motions (e) Residual images after motion compensation with each homography motion parameters

Figure 3: Extracted homography groups