CONSTRAINED, GLOBALLY OPTIMAL, MULTI-FRAME MOTION ESTIMATION

Sina Farsiu[1], **Michael Elad** [2], **Peyman Milanfar** [3]

[1,3] Electrical Engineering Department, University of California Santa Cruz
Email: {farsiu, milanfar}@ee.ucsc.edu
[2] Computer Science Department, The Technion, Israel Institute of Technology
Email: elad@cs.technion.ac.il

ABSTRACT

We address the problem of estimating the relative motion between the frames of a video sequence. In comparison with the commonly applied pairwise image registration methods, we consider global consistency conditions for the overall multi-frame motion estimation problem, which is more accurate. We review the recent work on this subject and propose an optimal framework, which can apply the consistency conditions as both hard constraints in the estimation problem, or as soft constraints in the form of stochastic (Bayesian) priors. The framework is applicable to virtually any motion model and enables us to develop a robust approach, which is resilient against the effects of outliers and noise. The effectiveness of the proposed approach is confirmed by a super-resolution application on synthetic and real data sets.

1. INTRODUCTION

Motion estimation with subpixel accuracy is of great importance to many image processing and computer vision applications, such as mosaicing [1] and super-resolution [1, 2]. Numerous image registration techniques have been developed throughout the years [3]. Of these, optical flow [4] [5], and correlation-based methods [6] are among the most popular.

These methods are mainly developed to estimate the relative motion between *a pair* of frames. For the cases where several images are to be registered with respect to each other (e.g. super-resolution applications), two simple strategies are commonly used. The first is to register all frames with respect to a single reference frame [7]. This may be called the *anchoring* approach. The choice of a reference or anchor frame is rather arbitrary, and can have a severe effect on the overall accuracy of the resulting estimates. This caveat aside, overall, this strategy is effective in cases where the camera motion is small and random (e.g. small vibrations of a gazing camera).

The other popular strategy is the *progressive* registration method [8], where images in the sequence are registered in pairs, with one of the pair acting as the reference frame. For instance, taking a causal view with increasing index denoting time, the i^{th} frame of the sequence is registered with respect to the $(i+1)^{\text{th}}$ frame and the $(i+1)^{\text{th}}$ frame is registered with respect to the $(i+2)^{\text{th}}$ frame, and so on. The motion between an arbitrary pair of frames is computed as the combined motion of the above incremental estimates. This method works best when the camera motion is smooth. However, in this method, the registration error between two "nearby" frames is accumulated and propagated when such values are used to compute motion between "far away" frames.

Neither of the above approaches take advantage of the important prior information available for the multi-frame motion estimation problem. This prior information constrains the estimated motion vector fields between any pair of frames to lie in a space whose geometry and structure, as we shall see in the next section, is conveniently described. In this paper, we study such priors and propose an optimal method for exploiting them, to achieve very accurate estimation of the relative motion in a sequence.

This paper is organized as follows. Section 2 introduces the consistency constraints in an image sequence and reviews the previous work on this subject. Section 3 describes the main contribution of this paper, which is an optimal framework for exploiting these consistency constraints. Using such framework, we introduce a highly accurate robust multi-frame motion estimation method, which is resilient to the outliers in an image sequence. Simulations on both real and synthetic data sequences are presented in Section 4, and Section 5 concludes this paper.

THIS WORK WAS SUPPORTED IN PART BY THE NATIONAL SCIENCE FOUNDATION GRANT CCR-9984246, US AIR FORCE GRANT F49620-03-1-0387, AND BY THE NATIONAL SCIENCE FOUNDATION SCIENCE AND TECHNOLOGY CENTER FOR ADAPTIVE OPTICS, MANAGED BY THE UNIVERSITY OF CALIFORNIA AT SANTA CRUZ UNDER COOPERATIVE AGREEMENT NO. AST-9876783. M. ELADS WORK WAS SUPPORTED IN PART BY JEWISH COMMUNITIES OF GERMANY RESEARCH FUND.

2. CONSTRAINED MOTION ESTIMATION

To begin, let us define $\mathbf{F_{i,j}}$ as the operator which maps (registers) frames indexed i and j as follows:

$$\underline{X}_i = \mathbf{F_{i,j}}(\underline{X}_j),$$

where $\underline{X}_i$ and $\underline{X}_j$ are the lexicographic reordered vector representations of frames i and j.

Now given a sequence of N frames, precisely $N(N-1)$ such operators can be considered. Regardless of considerations related to noise, sampling, and the finite dimensions of the data, there are inherent intuitive relationships between these pair-wise registration operators. In particular, the first condition dictates that the operator describing the motion between any pair of frames must be the composition of the operators between two other pairs of frames. More specifically, as illustrated in Figure 1(a), taking any triplet of frames i, j, and k, we have the first motion consistency condition as:

$$\forall i, j, k \in \{1, ..., N\}, \quad \mathbf{F_{i,k}} = \mathbf{F_{i,j}} \circ \mathbf{F_{j,k}}. \qquad (1)$$

The second rather obvious (but hardly ever used) consistency condition states that the composition of the operator mapping frame i to j with the operator mapping frame j to i should yield the identity operator. This is illustrated in Figure 1(b). Put another way,

$$\forall i, j \in \{1, ..., N\}, \quad \mathbf{F_{j,i}} = \mathbf{F_{i,j}^{-1}}. \qquad (2)$$

These natural conditions define an algebraic group structure (a Lie algebra) in which the operators reside. Therefore, any estimation of motion between frames of a ($N \gg 2$) image sequence could take these conditions into account. In particular, the optimal motion estimation strategy can be described as an estimation problem over a group structure, which has been studied before in other contexts [9].

The above properties describe what is known as the Jacobi condition, and the skew anti-symmetry relations [10]. For some practical motion models (e.g. constant motion or the affine model), the relevant operators could be further simplified. For example, in the case of translational (constant) motion, the above conditions can be described by simple linear equations relating the (single) motion vectors between the frames:

$$\forall i, j, k \in \{1, ..., N\}, \quad \underline{V}_{i,k} = \underline{V}_{i,j} + \underline{V}_{j,k}, \qquad (3)$$

where $\underline{V}_{i,j}$ is the motion vector field between the frames i and j. Note that $\underline{V}_{i,i} = \underline{0}$, and therefore the skew anti-symmetry condition is represented by (3), when $k = i$.

For the sake of completeness, we should note that the above ideas have been already studied to some extent in the computer vision community. In particular, the Bundle Adjustment (BA) [11] technique is a general, yet computationally expensive method for producing a jointly optimal 3D structure and viewing parameters, which bares close resemblance to what is proposed here. It is important to note that BA is not intended for motion estimation in 2-D images, and does not specifically take the algebraic group structure into account. Instead, it relies on an iterative method, which is largely based on the motivating 3-D application. On another front, to solve mosaicing problems, [12] adapted the BA method to a 2-D framework, where the estimated motion vectors are refined in a feedback loop, penalizing the global inconsistencies between frames. Also, the importance of consistent motion estimation for the Super-Resolution problem is discussed in [5].

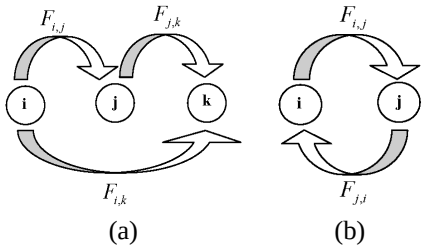


Fig. 1. The consistent flow properties: (a) Jacobi Identity and (b) Skew Anti-Symmetry.

The very recent approach in [10] exploits the Lie Group structure indirectly. The motions are estimated in an unconstrained framework, then "projected" to the set of valid motions by what the author calls Lie-algebraic averaging. While the framework of this approach is close to what we suggest, the algorithm presented there is suboptimal in that it uses the constraints only as a mechanism for post-processing already-estimated motion fields, resulting in a suboptimal overall procedure. Finally, in a similar way, another recent paper, [13], computes the motion vectors between a new frame and a set of frames for which relative motion vectors has been previously computed. Then, the motion vectors computed for the new image are used to refine the pairwise estimated motion of other frames. This two-step algorithm is iterated until convergence.

The framework we propose in this paper unifies the earlier approaches and presents an optimal framework where the constraints are used directly in the solution of the problem, and not simply as a space onto which the estimates are projected.

3. PRECISE ESTIMATION OF TRANSLATIONAL MOTION WITH CONSTRAINTS

We now describe our proposed methodology, and compare it against two other competing approaches. To simplify the

notation, we define the vectors $\underline{X}$, $\underline{V}$, and $\underline{V}(i)$ as follows:

$$\underline{X}=\begin{bmatrix}\underline{X}(1)\\ \underline{X}(2)\\ \vdots\\ \underline{X}(N)\end{bmatrix},\underline{V}=\begin{bmatrix}\underline{V}(1)\\ \underline{V}(2)\\ \vdots\\ \underline{V}(N)\end{bmatrix},\underline{V}(i)=\begin{bmatrix}\underline{V}_{i,1}\\ \vdots\\ \underline{V}_{i,j(i\neq j)}\\ \vdots\\ \underline{V}_{i,N}\end{bmatrix}, \quad (4)$$

where $\underline{X}(i)$ is the i^{th} image in this sequence rearranged in the lexicographic order. The vector $\underline{V}(i)$ contains the set of motion vector fields computed with respect to the reference frame i.

3.1. Optimal Constrained Multi-Frame Registration

In a general setting, the optimal solution to the multi-frame registration problem can be obtained by minimizing the following cost function:

$$\widehat{\underline{V}} = \underset{\underline{V}}{\text{ArgMin}}\, \Phi(\underline{X},\underline{V}) \text{ such that } \Psi(\underline{V})=0, \quad (5)$$

where Φ represents a motion-related cost function (e.g. penalizing deviation from brightness constraint, or a phase-based penalty), and Ψ captures the constraints discussed earlier.

To get a feeling for this general formulation, we address the translational motion case, (the consistency conditions for the affine case are described in the Appendix), with Φ representing the Optical Flow model:

$$\Phi(\underline{X},\underline{V})=\underbrace{\sum_{i,j=1}^{N}}_{i\neq j}\|\underline{X}_i^{(x)}v_{i,j}^{(x)}+\underline{X}_i^{(y)}v_{i,j}^{(y)}+\underline{X}_{i,j}^{(t)})\|_2^2, \quad (6)$$

where $\underline{X}_i^{(x)}$ and $\underline{X}_i^{(y)}$ are the spatial derivatives (in x and y directions) of the i^{th} frame, and $\underline{X}_{i,j}^{(t)}$ is the temporal derivative (e.g., the difference between frames i and j). Here the motion vector field $\underline{V}_{i,j}$ is spatially constant, and it can be represented by the scalar components $v_{i,j}^{(x)}$ and $v_{i,j}^{(y)}$ in x and y axes, respectively, and for $1 \leq i,j \leq N$. Using this, the translational consistency condition as in Equation (3) is then formulated as

$$\Psi(\underline{V}): \quad C\underline{V}=0, \quad (7)$$

where the unknown motion vector $\underline{V}$ has all the $2N(N-1)$ entries $v_{i,j}^{(x)}$ and $v_{i,j}^{(y)}$ stacked to a vector. The constraint matrix C is of size $[2(N-1)^2 \times 2N(N-1)]$. Each row in C has only two or three non-zero (± 1) elements representing the skew anti-symmetry and Jacobi identity conditions in (3), respectively. The defined problem has a quadratic programming structure, and it can be solved using accessible optimization algorithms.

3.2. Two-Step Projective Multi-Frame Registration

As a comparison to our proposal, we discuss a two-step approach that is in spirit similar to what is done in [10]. In this method, for a sequence of N frames, in the first step all $N(N-1)$ possible pairwise motion vector fields ($\underline{V}_{i,j}$) are estimated. Note that the pairwise motion vector fields are individually estimated by optimizing the following cost function:

$$\widehat{\underline{V}}_{i,j} = \underset{\underline{V}_{i,j}}{\text{ArgMin}}\, \Phi(\underline{X}_{i,j},\underline{V}_{i,j})$$

where $\Phi(\underline{X}_{i,j},\underline{V}_{i,j})$ may represent any motion estimation cost function.

In the second step, these motion vectors are projected onto a consistent set of $N(N-1)$ pairwise motion vectors. For the case of translational motion model, with the consistency condition in (7), the projection of the motion vector fields onto the constraint space is computed as:

$$\underline{V}_{proj} = (I - C[C^TC]^{-1}C^T)\widehat{\underline{V}}.$$

Such a two step projection method (as in [10]) is not optimal and would be expected to result in inferior estimations compared to the solution of the method posed in Equation (5).

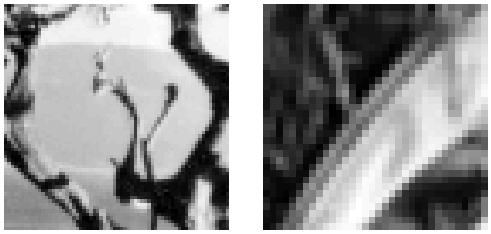
3.3. Robust Multi-Frame Registration

In many real video sequences, the practical scenarios are not well-modelled by temporally stationary noise statistics, and abrupt changes or occlusions may introduce significant outliers into the data. Note that even the presence of very small amount of outliers, which may be unavoidable (e.g. the bordering pixel effects), heavily affects the motion estimation accuracy. In such cases, it is prudent to modify the above approaches in two ways. First, one may replace the hard constraints developed above with soft ones, by introducing them as Bayesian priors which will *penalize* rather than *constrain* the optimization problem. Second, we may want to introduce alternative norms to the standard 2-norm for both the error term and the constraint in (5). Incorporating both modifications, one can consider optimizing a modified cost function which includes a term representing the "soft" version of the constraints as:

$$\widehat{\underline{V}} = \underset{\underline{V}}{\text{ArgMin}}\, \Phi_r(\underline{X},\underline{V}) + \lambda\Psi(\underline{V}), \quad (8)$$

where λ represents the strength of the regularizing term. The functions Φ and Ψ may use robust measures, such as the 1-norm. For instance, to deal with outliers directly, one might use

$$\Phi_r(\underline{X},\underline{V})=\underbrace{\sum_{i,j=1}^{N}}_{i\neq j}\|\underline{X}_i^{(x)}v_{i,j}^{(x)}+\underline{X}_i^{(y)}v_{i,j}^{(y)}+\underline{X}_{i,j}^{(t)})\|_1. \quad (9)$$



a: Experiment 1 b:Experiment 2

Fig. 2. One of the input frames used in the first and the second experiments (simulated motion).

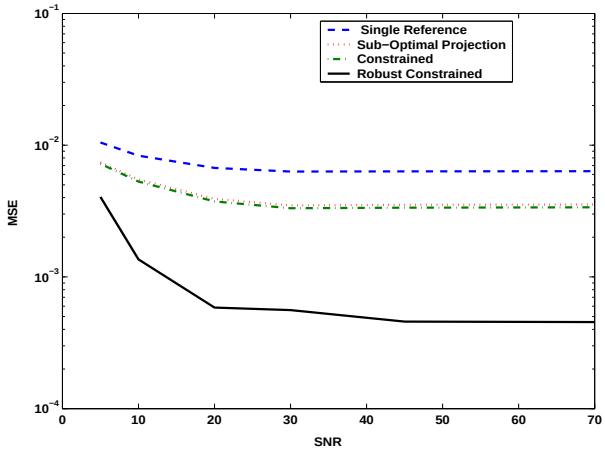


Fig. 3. MSE comparison of different registration methods in the first simulated experiment.

The use of such robust error term together with the hard constraint cost function of (5) often suffices to enhance the estimator performance. Note that unlike the L_2 norm which reduces the estimation error by an implicit averaging of the estimates, the robust L_1 norm implements a median estimator [14], which effectively picks the most reliable estimated motion vector for each pair of frames. The experiments in the next section justify this claim.

4. EXPERIMENTS

A simulated experiment was conducted by registering 5 frames of size $[65 \times 65]$. For these frames we have the correct translational motion vectors in hand. One of these frames is shown in Fig.2(a).

The mean square errors (MSEs) of the computed motion vectors (against the true motion vectors) with the single reference approach (Section 1), suboptimal projective (Section 3.2), the L_2 constrained (Section 3.1), and the L_1 norm with hard constraints (Section 3.3) methods are compared in Fig.3. Each point in this graphs shows the average of 100 different realizations of additive noise (Monte Carlo simulation) for different SNRs [1].

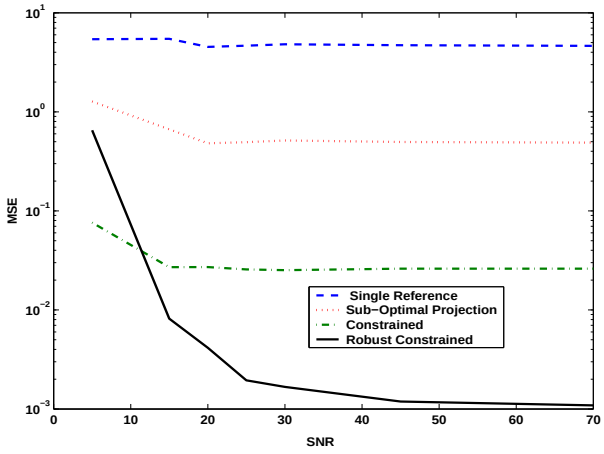


Fig. 4. MSE comparison of different registration methods in the second simulated experiment.

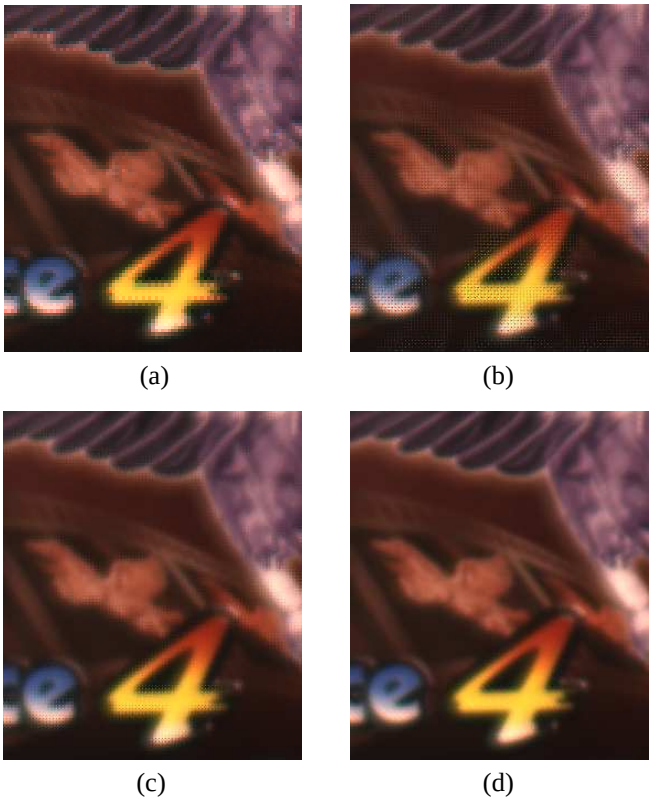
The second simulated experiment was conducted by registering 7 frames of size $[39 \times 39]$. One of these frames is shown in Fig. 2(b). We repeated the previous experiment on this data set (with 30 Monte Carlo iterations for different SNRs) and compared the performance of different methods in Fig. 4.

A real experiment was also conducted aligning 27 color-filtered low-resolution (LR) images. One of these LR frames after demosaicing [2] is shown in Fig.5(a). The method of [2] was used to construct a high resolution (HR) image, by registering these images on a finer grid (resolution enhancement factor of three in x and y directions). We used the method described in [15] to compute the motion vectors in an "anchored" fashion (Section 1). Figure 5(b) shows the HR reconstruction using this method with clear mis-registration errors. The result of applying the two step multi-frame projective image registration of Section 3.2 is shown in Fig.5(c). Some mis-registration errors are still visible in this result. Finally, the result of applying the optimal multi-frame registration method (Section 3.1) is shown in Fig.5(d), with almost no visible mis-registration error.

5. CONCLUSION

In this paper we studied several multi-frame motion estimation methods, focusing on the methods that exploit the consistency conditions. As an alternative to existing methods, we proposed a general framework to optimally benefit

[1] Signal to noise ratio (SNR) is defined as $10 \log_{10} \frac{\sigma^2}{\sigma_n^2}$, where σ^2, σ_n^2 are variance of a clean frame and noise, respectively.



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

Fig. 5. Experimental registration results for a real sequence. (a) One input LR frame after demosaicing.(b) Single Reference HR registration. (c) Projective HR registration. (d) Optimal HR registration. This paper (with all color pictures) is available at *http://www.soe.ucsc.edu/∼milanfar*.

from these constraints. Such framework is flexible, and is applicable to more general motion models. Based on this framework, we proposed a highly accurate multi-frame motion estimation method which is robust to the outliers in image sequences. This robust method, which minimizes an L_1 norm cost function, often provides more accurate estimation than the common least square approaches. Our experiments show that the high accuracy and reliability of the proposed multi-frame motion estimation method is especially useful for the multi-frame super-resolution applications in which very accurate motion estimation is essential for effective image reconstruction.

6. ACKNOWLEDGMENTS

We would like to thank Eyal Gordon from the Technion-Israel Institute of Technology for helping us capture the raw CFA images used in the Fig.5 experiment.

Appendix: Affine Motion Constraints

In this section we review the consistency constraints for the affine motion model. The affine transformation models a composition of rotation, translation, scaling, and shearing. This six parameter global motion model is defined by

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \begin{bmatrix} a_{i,j} & b_{i,j} \\ c_{i,j} & d_{i,j} \end{bmatrix} \begin{bmatrix} x_j \\ y_j \end{bmatrix} + \begin{bmatrix} e_{i,j} \\ f_{i,j} \end{bmatrix}, \quad (10)$$

where $[x_i, y_i]^T$, and $[x_j, y_j]^T$ are the coordinates of two corresponding pixels in frames i and j. Defining

$$M_{i,j} = \begin{bmatrix} a_{i,j} & b_{i,j} \\ c_{i,j} & d_{i,j} \end{bmatrix}, \quad \underline{T}_{i,j} = \begin{bmatrix} e_{i,j} \\ f_{i,j} \end{bmatrix}, \quad (11)$$

the consistency constraints for the affine case are defined by the relations

$$\forall \quad 1 \le i,j,k \le N, \quad \begin{cases} M_{i,k} &= M_{i,j} M_{j,k} \\ \underline{T}_{i,k} &= M_{i,j} \underline{T}_{j,k} + \underline{T}_{i,j} \end{cases} \quad (12)$$

Note that $M_{i,i} = I$ and $\underline{T}_{i,i} = \underline{0}$, and therefore (12) results in a set of $6(N-1)^2$ independent nonlinear constraints.

A more intuitive (and perhaps more practical) set of constrains can be obtained if we consider a simplified version of the general affine model where only scale, rotation, and translation are considered. Such simplified model is represented by replacing the first coefficient matrix on the right side of (10) with

$$M'_{i,j} = \begin{bmatrix} a_{i,j} & b_{i,j} \\ c_{i,j} & d_{i,j} \end{bmatrix} = \alpha_{i,j} \begin{bmatrix} \cos(\theta_{i,j}) & -\sin(\theta_{i,j}) \\ \sin(\theta_{i,j}) & \cos(\theta_{i,j}) \end{bmatrix} \quad (13)$$

where $\alpha_{i,j}$, and $\theta_{i,j}$ are the scaling and rotation parameters, respectively. The consistency constraints for this simplified affine model are given by the following relations:

$$\begin{cases} \alpha_{i,k} &= \alpha_{i,j} \alpha_{j,k} \\ \theta_{i,k} &= \theta_{i,j} + \theta_{j,k} \\ \underline{T}_{i,k} &= M'_{i,j} \underline{T}_{j,k} + \underline{T}_{i,j} \end{cases} \quad (14)$$

For a set of N frames the above relations amount to $4(N-1)^2$ independent non-linear constraints. Non-linear programming (e.g. "*fmincon*" function in MATLAB) can be used to minimize the cost functions with such non-linear constraints.

7. REFERENCES

[1] A. Zomet and S. Peleg, "Efficient super-resolution and applications to mosaics," in *In Proc. of the Int. Conf. on Pattern Recognition (ICPR)*, Sept. 2000, pp. 579–583.

[2] S. Farsiu, M. Elad, and P. Milanfar, "Multi-frame demosaicing and super-resolution of color images," *to appear in IEEE Trans. Image Processing*, 2005.

[3] L.G. Brown, "A survey of image registration techniques,," *ACM Computing Surveys*, vol. 24, no. 4, pp. 325–376,, 1992.

[4] B. Lucas and T. Kanade, "An iterative image registration technique with an application to sterio vision," in *In Proc. of DARPA Image Understanding Workshop*, 1981, pp. 121–130.

[5] W.Y. Zhao and H.S. Sawhney, "Is super-resolution with optical flow feasible?," in *ECCV*, 2002, vol. 1, pp. 599–613.

[6] M. Alkhanhal, D. Turaga, and T. Chen, "Correlation based search algorithms for motion estimation," in *Picture Coding Symposium*, Apr. 1999.

[7] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multi-frame super-resolution," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[8] L. Teodosio and W. Bender, "Salient video stills: Content and context preserved," in *In Proc. of the First ACM Int. Conf. on Multimedia*, Aug. 1993, vol. 10, pp. 39–46.

[9] S. Marcus and A. Willsky, "Algebraic structure and finite dimensional nonlinear estimation," *SIAM J. Math. Anal.*, pp. 312–327, Apr. 1978.

[10] V.M. Govindu, "Lie-algebraic averaging for globally consistent motion estimation," in *In Proc. of the Int. Conf. on Computer Vision and Patern Recognition (CVPR)*, July 2004, vol. 1, pp. 684–691.

[11] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment – a modern synthesis," in *Vision Algorithms: Theory and Practice*, 2000, vol. 1883 of *Lecture Notes in Computer Science*, pp. 298–372.

[12] H.S. Sawhney, S.C. Hsu, and R. Kumar, "Robust video mosaicing through topology inference and local to global alignment," in *ECCV*, 1998, vol. 2, pp. 103–119.

[13] Y. Sheikh, Y. Zhai, and M. Shah, "An accumulative framework for the alignment of an image sequence," in *ACCV*, Jan. 2004.

[14] P. J. Huber, *Robust Statitics*, Wiley, New York, 1981.

[15] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierachical model-based motion estimation," *In Proc. of the European Conf. on Computer Vision*, pp. 237–252, May 1992.