Robust multi-body segmentation

Andreas Dante, Mike Brookes, A. G. Constantinides
Dept. of E.E.E., Imperial College London,
London SW7 2BT, UK
{andante,mike.brookes,a.constantinides}@imperial.ac.uk

Abstract

Good correspondences are a key to a correct 3D reconstruction of a scene, especially in the presence of multiple independent objects. In this paper a novel segmentation algorithm is presented that decouples the outlier rejection from the object segmentation. It is shown that the proposed outlier rejection scheme provides a dense set of correspondences across the image and eliminates gross outliers. These correspondences are subsequently used for the segmentation of objects by enforcing constraints of rigid motion. Simple additional constraints are incorporated in the segmentation process to ensure further stability under non-optimal conditions. The algorithm requires only two pictures of the scene and most of the computation can be parallelised easily, making the algorithm highly suitable for real-time hardware processing. The performance of the algorithm is illustrated on real images and the strengths and weaknesses of the approach are discussed.

1 Introduction

The detection of independently moving rigid objects is currently a major challenge in computer vision. Once solved it allows the 3D reconstruction of a general motion scene containing multiple rigid objects.

The rigid motion constraint imposed by epipolar geometry has been widely used to segment objects [7, 13, 18]. In many approaches however the same constraint has also been applied to reject correspondence outliers [7, 13]. Experiments have shown that using the same constraint for both targets reduces the stability of the segmentation and reconstruction.

In this paper a novel method is presented that allows the separation of outlier rejection and object segmentation into two distinct processes using independent constraints. Outlier rejection is performed using vector field constraints and hence the rigid motion constraint can be used exclusively for the segmentation of objects. To reduce the complexity of the vector field constraints, a rectangular grid of initial correspondences is obtained by means of block matching. Since block matching also provides correspondences for untextured areas, a large number of outliers occur, which need to be filtered. The purpose of this publication is to demonstrate how the gross outliers of the block correspondences can be removed efficiently such that independent objects can be segmented robustly on the basis of rigid motion constraints.

The following contributions are made: 1. A new concept of constraint separation for multi-body segmentation is introduced. 2. A method for refining the precision of block

BMVC 2003 doi:10.5244/C.17.68

correspondences is presented. 3. An extended method for the rejection of outliers from block correspondences is developed. 4. A segmentation algorithm is constructed that incorporates the preceding methods.

The paper starts with a review of the literature followed by an introduction to block matching. Different methods for the improvement of correspondences are then introduced and additional steps for the segmentation are described. The paper concludes with a summary of the complete segmentation algorithm, a presentation and a discussion of results.

2 Background

2.1 Correspondence Task

Two main methods have been applied to the correspondence task: the feature-extraction and matching based approach and the optical-flow-based approach. Both are reviewed and compared in [1, 9, 17]. Apart from these methods, block matching has been rarely used in computer vision [3, 5, 11]. The main drawbacks of block matching are that it delivers correspondences even in areas where the reliability is low, that it does not account for rotations or perspectives on the block level and that features cannot be tracked in the same way as salient features. The benefits are a regular grid of correspondences, a wide spread of features across the image (and thus the potential of having sufficient features for all objects), and the availability of hardware implementations [20].

2.2 Outlier removal

Outlier rejection has been well researched for salient feature correspondences: [6, 13, 15, 16]. Since block matching has mainly been applied to video compression techniques such as MPEG-2, where the underlying true projected motion is irrelevant, few outlier rejection schemes have been developed for block matching. One example is the vector median filter [2], which has also been applied in the field of computer vision to track objects [11]. This however does not take into account the characteristics of the vector field resulting from the perspective projection of 3D motion.

2.3 Multi-body segmentation

An early approach of robust motion segmentation was published in [15] followed by extensions in [7, 13, 16]. In these approaches, outliers of extracted and matched feature points are removed using the RANSAC algorithm [6]. Objects are then clustered into different motion models using for example a Bayesian approach on the basis of the constraints enforced by rigid motion. The problem with these approaches is that outliers of one motion are potential inliers for another motion. If a scene contains a large proportion of outliers it is hard to distinguish outliers from additional objects.

A robust approach based on the factorization method is presented in [10, 19]. The shape interaction matrix of the factorization method is rearranged to block-diagonal submatrices representing the independently moving objects.

The concept of a Multi-body Fundamental Matrix was introduced in [18]. The motion models of the objects are obtained by transferral of the multi-body constraint to a higher dimensional space and subsequent polynomial factorization.

In contrast to many of the concepts in the literature, the approach of this paper aims to decouple the outlier correspondence problem from the problem of the motion detection of independent objects. The proposed method is based on block vector fields and therefore provides vectors for most objects of the scene, which is not always the case for methods based on salient feature point matching.

3 Obtaining motion vectors

3.1 Block Matching

In block matching the initial image is divided into rectangular shaped blocks. The best match for each block in the initial image is searched for in the subsequent image. The 2-dimensional search is usually limited by a window, and the search steps within that window can differ depending on the precision required (e.g. full or half-pixel). The *best match* has the least absolute or squared difference of luminance values summed over all block pixels. The cross correlation may be used as an alternative criterion of best match. The error across the 2-dimensional search window is referred to as *error surface* in the remainder of this paper.

As presented in [4], a sub-pixel precision can be obtained from block-matching even if block matching is only performed on a full-pixel image. For this it is assumed that the error surface is monotonic towards the minimum in the 1-pixel neighbourhood of the best match. The method uses a least-squares criterion to fit a quadratic surface to the nine points in the direct vicinity of the full pixel best match.

3.2 Improving the block-matching result

In order to derive the correct motion of objects, many precise motion vectors are required. Block matching is often implemented such that the motion vector assigned to each block is the position of the least squared difference of luminance values over all block pixels. While this *best match* represents the true motion on the image plane of well textured blocks, it is not necessarily the correct motion if the texture is weak or if there are large perspective distortions between the two views.

An in-depth analysis of error surfaces of real image sequences has been carried out. For the sake of a simpler representation, the reciprocal of the error, referred to as *match quality,* has been analyzed. The position of the maximum of the match quality coincides with the minimum of the error surface. A typical match quality surface of a region with perspective distortion and weak texture is show in Figure 1.

The true motion of the block shown in Figure 1 is represented by the sharp peak “(1)”. This however is not the maximum match quality “(2)”. In this publication the expression *sharp peak* is used for a pixel whose match quality is of at least two thirds of the maximum match quality, larger than that of all its neighbouring pixels and by a factor of 1.5 larger than the pixels with distance of two pixels in a circle around the peak. This heuristic definition takes into account that the peak may be in a sub-pixel position between two pixels and thus it may be blurred in a region of adjacent pixels.

Beside the maximum match quality, the tallest sharp peak (if it exists) is transferred to the following stage, where one of the two is filtered depending on vector field constraints.

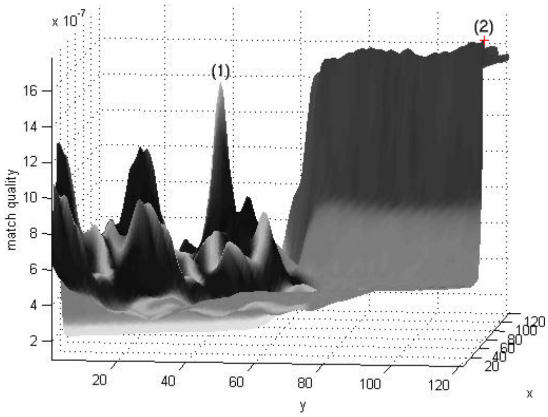


Figure 1: Match quality for a block with little texture and perspective distortion

Block matching uses a shift of the block to find the best match. At straight line edges in the image this results in several similar "good match" peaks along the edge. Experiments have shown that due to quantisation, the correct vector often has a smaller peak than several of the incorrect peaks. Since not all correspondences are needed, the vector of a block on an edge within a weakly textured area is removed due to unreliability. A problematic edge is detected by creating a set of pixels with match quality above a threshold (for example 1/5 of the maximum). If the points in the set have a wide spread (for example 10 pixels) in at least one direction, and a straight line can fit to these points, such that the mean distance of all points to the line is below a threshold (for example 5 pixels), then the block is rejected as a potential outlier.

To further enlarge the block matching vector fields under difficult light conditions, the vectors obtained through two implementations of block matching are combined. In one implementation the mean luminance of the block is subtracted before obtaining the summed squared difference; in the other the values are left un-normalized. This provides a larger set of correct correspondence vectors. Tests have shown that the cross-correlation in general does not provide better results than the normalized summed squared difference.

4 Vector field constraints

The motion vector fields on the image plane are governed by the 3D motion model and the projection model of a general scene.

An analysis of motion vector fields [4] for locally planar objects has shown that, in the absence of occlusions, they vary smoothly across the image plane regardless of object motion, camera motion and focal length changes. Furthermore, the relative change between neighbouring motion vectors is small except in the vicinity of a null vector. Since this applies even to the case of multiple large objects with independent motion, it allows the filtering of vector fields on the basis of the change of vectors relative to its neighbouring blocks.

Three criteria for filtering motion vector fields are described in the following. These incorporate the above mentioned constraints on the motion vector field. Two of the criteria have been proposed in [4] and are referred to as *neighbourhood* and *smoothness* criterion respectively. The third is called *clipping* criterion and is proposed in this publication. The neighbourhood criterion (C_1) accepts blocks in regions with uniform motion vectors and the smoothness criterion (C_2) accepts blocks in regions where motion vectors are changing linearly over the image. The *clipping* criterion C_3 accounts for luminance clipping of the image sensor, which is a common problem found in small areas of many available test images. In the event of illumination changes, several block vectors may point to the same pixel region; in particular when an image region with clipped values is larger in the first image than in the second image. In this case several blocks in the first image match with the same image block in the second image. Vectors of these blocks are filtered by the neighbourhood criterion but not filtered by the smoothness criterion. Therefore a combination of the smoothness criterion with the clipping criterion is proposed below (2).

The three constraints can be described by the following formulae:

$$\begin{aligned} C_1 &= \Big(\sum_{i}^{8Neighb.} (\|v_i - v_0\| \leq K_1 \cdot \|v_0\|)\Big) \geq T_1 \\ C_2 &= \Big(\sum_{i}^{8Neighb.} (\|\tfrac{v_i + v_{-i}}{2} - v_0\| \leq K_2 \cdot \|v_0\|)\Big) \geq T_2 \\ C_3 &= \Big(\sum_{i}^{8Neighb.} (\|v_i + \beta_i - v_0\| \leq K_2 \cdot \|v_0\|)\Big) \leq T_2 \end{aligned} \tag{1}$$

where $\|a\|$ is the Euclidean length of vector a and v_0 is the vector of the block under consideration. v_i represents the vectors of one of the eight neighbouring blocks and v_{-i} is the neighbouring block opposite of v_i with respect to the block under consideration. The inequalities return 1 if true and 0 otherwise, allowing them to be summed. β_i denotes the position in pixels of the block of v_i relative to the block of v_0. K_1, K_2, T_1, T_2 are threshold values. In [4] the parameter values: $K_1 = 8\%$, $K_2 = 3\%$, $T_1 = 3$, $T_2 = 4$ were determined to be good for a wide range of scenes.

A block vector is accepted as inlier if the following expression is true:

$$C_1 \;\; OR \;\; (C_2 \;\; AND \;\; C_3) \tag{2}$$

Blocks along the image border are rejected by this method. This uncontrolled rejection can be avoided by applying the following scheme: A border block is accepted if the criterion (2) is fulfilled with a stricter set of parameters (T_1 and T_2) for the closest adjacent block towards the center of the image. The parameters T_1 and T_2 are increased in this context to account for the limitation of the vector field and the large scale occlusions at the image borders.

In the presence of multiple possible motion vectors per block, as obtained from the block matching stage (Section 3.2), the algorithm needs to be extended. For each of the two best matching motion vectors of each block in the eight-neighbourhood of a reference block, the vector field constraints (2) are applied. Since nine blocks are used for the analysis, and a large proportion of blocks has only a single vector, the chance of accepting an outlier vector is much reduced.

5 Rigid motion constraint

Once a reliable set of correspondences is obtained RANSAC [12] can be used to obtain the Fundamental Matrix (or another motion model) of the primary object (e.g. the background). RANSAC incorporates the epipolar constraint of rigidity of an object. Therefore the correspondences of independently moving objects will be labeled as outliers of the first object. Applying RANSAC again to the set of outliers of the first object can identify the next most dominant object and so forth for all other objects until the set of correspondences is too small to obtain another motion model.

If the set of correspondences contains many outliers, RANSAC may randomly select a set of correspondence outliers and interpret it as an independently moving object. This results in a wrong segmentation and is the justification for the extensive filtering described in Section 4.

6 Colour Grouping

The motion vectors of blocks with intensive texture are very reliable, so long as the texture is not repetitive. Areas with little texture in general do not contain reliable motion information. However the colour of these areas is often homogenous; this assumption is used for the presented segmentation algorithm: A pixel in the image is grouped to the same object as a local correspondence pixel if it fulfills two criteria: 1) Its colour is within a threshold in the same region of the colour of the correspondence block center pixel; 2) There is a path of pixels satisfying criterion 1 between the correspondence block center pixel and the examined pixel (locality constraint).

7 Algorithm

The complete algorithm for robust multi-body segmentation is laid out below:

1. **Motion vectors:** Obtain the enhanced motion vector field for the image as described in Section 3.

2. **Outlier Rejection:** Filter the motion vectors on the basis of vector field constraints as described in Section 4.

3. **Rigidity Constraint:** (a) Apply RANSAC [12] to the correspondence vector field to pick the correspondences of the primary object and to obtain its Fundamental Matrix (or homography if planar). (b) Determine the residual error of all correspondences and group those that are further than a threshold to belong to separate objects. Repeat steps (a) and (b) for all objects until the number of correspondences is too small to establish a motion model (Section 5).

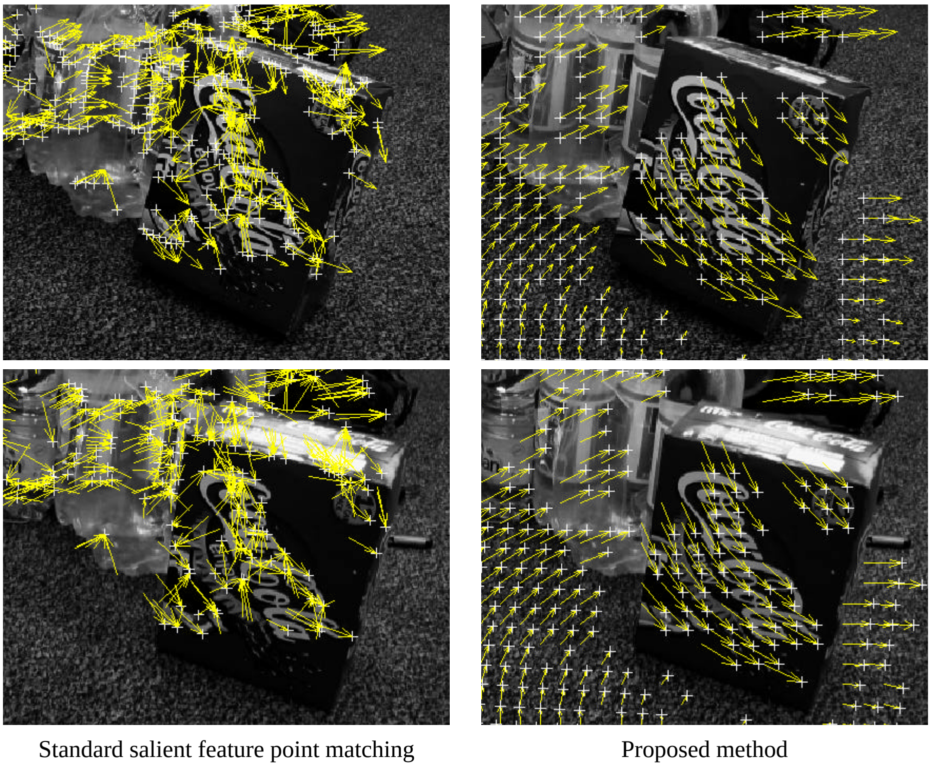
4. **Grouping by colour:** Group to each object those image pixels which are in the connected neighbourhood of similar colour values (Section 6).

8 Results

Results from real image sequences are shown in this section. In the first subsection the results of the correspondence retrieval (Sections 3 and 4) are presented and thereafter the multi-body segmentation (Section 7).

8.1 Obtaining correspondences

A scene with a rigidly moving object viewed by a rotating and translating camera has been chosen to demonstrate the properties of the algorithm on real image data. The result of the correspondence matching has been compared to the results of [14] (cross correlation of Harris corners) using the default parameters and an initial set of 1200 correspondences. It is stated in [14] that the methods used for the correspondence task are standard techniques and not the best available techniques.


Standard salient feature point matching

Proposed method

Figure 2: Correspondence results: A comparison between standard salient feature-based method [14] (left) and the proposed method for correspondence extraction (right). The top row shows a section of the first image and the bottom row shows the same section of the second image of the scene. The correspondences are superimposed.

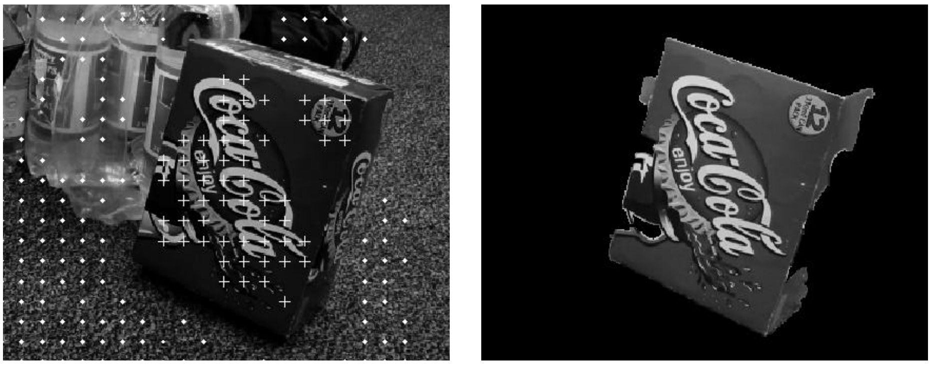
Figure 2 shows the results of the correspondence estimation from a section of the original image before (upper row) and after the camera and object motion (lower row of images). The two images on the left hand side depict the results of the salient feature extraction and matching method and the two images on the right hand side show the result of the proposed algorithm. The main motion of the camera is a translation to the right and an anticlockwise pan. The cola box moves independent of the background: It tilts forward and rotates anticlockwise by a small angle. Correspondence vectors and correspondences (shown as crosses) are superimposed on the images. In the ideal case, the cross at a vector tail in the upper image should cover the same image content as the cross at the vector tip in the bottom image.

It can be seen clearly that the salient feature matching method provides many incorrect correspondences and does not provide correspondences for a large portion of the image segment. In contrast to this, the proposed method provides a dense correspondence vector field with no visible outliers. The difference is largest for the highly structured carpet region, revealing a weakness of the salient feature extraction method.

The proposed algorithm has rejected 53% of the 1200 original blocks. 63 of the 1200 blocks (5.3%) have supplied two vectors to the filtering stage, of which one has been filtered. By means of visual inspection, none of the remaining 565 correspondence vectors has been found to be an outlier. Similar results have also been achieved on a variety of other sequences. The residual error [8] of all the background features with respect to the Epipolar geometry of the background is 0.58. The residual error of all the features on the cola box is 1.21 with respect to their own Epipolar geometry and 1444 with respect to the Epipolar geometry of the background. These low errors with respect to the own Epipolar geometry and large errors with respect to the Epipolar geometry of other objects allow a clear and stable segmentation of the scene.

8.2 Multi-body segmentation

The results of the proposed segmentation algorithm (Section 7) are demonstrated on the example of the scene described in Section 8.1. The image on the left hand side of Figure 3 shows the correspondences that have been grouped to the background as superimposed white dots and the features that have been grouped to the independently moving object as white crosses. The image on the right hand side shows how the grouping has been extended accross the object (cola box) by the methods described in Section 6.



First image with correspondences of the object (+) and background (.) superimposed

Segmentation result

Figure 3: Segmentation result using the proposed algorithm for a scene with an independently moving object and free-hand camera motion.

The independent motion has been detected correctly and most of the area of the cola container is segmented properly, demonstrating that the algorithm performs well in practical situations. Parts of the left and upper side of the box are not included due to occlusion and reflections in the second image (shown in the lower row of Figure 2).

9 Discussion

Rotations and perspective distortions on the block area are not taken into account by block matching. However, it was observed that only few block vectors are severely affected by this and that most of the affected blocks are filtered out by the algorithm. The vectors of the remaining blocks have relatively small errors compared to correspondence outliers. Due to the effective outlier rejection, least-squares techniques may be applied before using the vectors for 3D reconstruction. As a future extension, relaxation schemes may be applied to the vector field constraints in order to extend the segmentation to non-rigid objects.

10 Conclusion

A novel algorithm for the robust segmentation of multi-body motion scenes is presented in this publication. A core element of the approach is a method for the rejection of correspondence outliers, which is not based on the epipolar constraint. This allows the application of the epipolar constraint exclusively to the segmentation of rigid objects. It is demonstrated on real images that the proposed method for the retrieval of correspondences performs particularly well on textured surfaces where the standard salient feature approach performs particularly poorly. It is also shown that additional correspondence reliability information can be obtained from the error surface of the matching process. The stability of the segmentation is enhanced by providing a large set of correspondences that covers all objects of the scene. The performance of the segmentation is illustrated on real images.

References

[1] J. K. Aggarwal and N. Nandhankumar. On the computation of motion from sequences of images - A review. *Proc. of the IEEE*, 76(8):917–935, August 1988.

[2] L. Alparone, M. Barni, F. Bartolini, and V. Cappellini. Adaptively weighted vector-median filters for motion-fields smoothing. In *International Conference on Acoustic Speech and Signal Processing*, 1996.

[3] Yen-Kuang Chen, Yun-Ting Lin, and S.Y. Kung. A feature tracking algorithm using neighborhood relaxation with multi-candidate pre-screening. In *Proceedings, International Conference on Image Processing*, pages 513–516, 1996.

[4] A. Dante and M. Brookes. Precise real-time outlier removal from motion vector fields for 3d reconstruction. In *Proceedings, International Conference on Image Processing*, 2003.

[5] L. Falkenhagen. Block-based depth estimation from image triples with unrestricted camera setup. In *IEEE First Workshop on Multimedia Signal Processing*, pages 280–285, 1997.

[6] M. Fischler and R. Bolles. Random sampling consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Communications of ACM*, 24(6):381–395, 1981.

[7] A.W. Fitzgibbon and A. Zisserman. Multibody structure and motion: 3-d reconstruction of independently moving object. In *Proceedings of the European Conference on Computer Vision*, volume 2, 2000.

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[9] M. Irani and P. Anandan. All about direct methods. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Proc. of the Int. International Workshop on Vision Algorithms*, pages 267–277, 1999.

[10] K. Kanatani. Motion segmentation by subspace separation and model selection. In *Int. Conf. Computer Vision and Pattern Recogn.*, volume 2, pages 301–306, 2001.

[11] L. Di Stefano and E. Viarani. Vehicle detection and tracking using the block matching algorithm. In *Proc. of 3rd IMACS/IEEE Int'l Multiconference on Circuits, Systems, Communications and Computer*, pages 4491–4496, 1999.

[12] P. H. S. Torr. *Outlier Detection and Motion Segmentation*. PhD thesis, Univ. of Oxford, 1995.

[13] P. H. S. Torr. Geometric motion segmentation and model selection. In J. Lasenby, A. Zisserman, R. Cipolla, and H. Longuet-Higgins, editors, *Philosophical Transactions of the Royal Society A*, pages 1321–1340. Royal Society, 1998.

[14] P. H. S. Torr. A structure and motion toolkit in matlab. Technical Report MSR-TR-2002-56, Microsoft Research, 7 JJ Thomson Avenue, Cambridge, CB3 0FB,UK, http://research.microsoft.com/˜philtorr/, 2002.

[15] P. H. S. Torr and D. W. Murray. Outlier detection and motion segmentation. In *Sensor Fusion VI*, pages 432–443. SPIE volume 2059, 1993. Boston.

[16] P. H. S. Torr and A. Zisserman. Concerning bayesian motion segmentation, model averaging,matching and the trifocal tensor. In H. Burkharddt and B. Neumann, editors, *European Conf. on Computer Vision*, volume 1, pages 511–528. Springer, 1998.

[17] P. H. S. Torr and A Zisserman. Feature based methods for structure and motion estimation. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *International Workshop on Vision Algorithms*, pages 278–295, 1999.

[18] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. Segmentation of dynamic scenes from the multibody fundamental matrix. In *Workshop on Vision and Modeling of Dynamic Scenes, European Conference of Computer Vision*, May 2002.

[19] Y. Wu, Z. Zhang, T.S. Huang, and J.Y. Lin. Multibody grouping via orthogonal subspace decomposition. In *IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 252–257, 2001.

[20] Yuan-Hau Yeh and Chen-Yi Lee. Cost-effective VLSI architectures and buffer size optim. for full-search block matching algorithm. *IEEE Trans. on VLSI Systems*, 7(3):345–358, Sept 1999.