A New Structural Constraint and its Application in Wide Baseline Matching

X. Lu and R. Manduchi
University of California, Santa Cruz
{xylu,manduchi}soe.ucsc.edu

Abstract

We introduce a new structural constraint that can be used for matching points in image pairs taken from a wide baseline. No assumption is made about the geometry of the 3–D points or of surfaces in the scene, nor about the location or orientation of the cameras. This structural constraint can be used to reduce the search space when matching a number of feature points in two images, eliminating the risk of selecting a matching that is provably wrong because unrealizable.

1. Introduction

This work considers the problem of point matching across images of the same scene taken by two uncalibrated cameras. Typical applications include the recovery of the epipolar geometry of the camera system (which allows for dense stereo measurements) and structure from motion from a moving camera [10].

If the displacement between the cameras is small, the two images may look very similar, and methods based on local visual appearance, combined with a reduced search window, may work well. However, when the cameras are at arbitrary location and orientation, the problem becomes more challenging, for a number of reasons. First, one cannot use the location of a feature in one image as a basis for searching the corresponding point in the other image, because such point may be in a very different location. Second, a possibly large number of feature points in one image may not have a correspondence in the other image, due to limited field of view, occlusions, and missed detection by the feature operator. Third, corresponding points in the two images may have rather different visual appearance.

Lacking constraints on the location of matching features, and with purely visual similarity criteria proving unreliable (and hunted by ambiguity), the problem of point matching under wide baseline is daunting. The combinatorial explosion of the number of potential matchings adds to this meagre scenario. Consider a simple case with N feature points in one image having one–to–one (but unknown) correspondence with N points in the other image. Our problem is to find a *matching*, that is, a chain of one–to–one correspondences of the two point sets, that is optimal under some given criterion. Since the number of potential matching is $N!$, exhaustive search is out of the question.

Clearly, some structure is needed to make this problem tractable. For example, it is well known [10] that seven correspondences may already be enough to determine the epipolar geometry, thereby constraining the other correspondences. A similar observation is at the basis of the RANSAC algorithm [10], which found widespread acceptance as a tool for this type of problems. Unfortunately, even random sampling may fail if the probability of selecting a correct correspondence set for epipolar geometry reconstruction is small.

Indeed, there exists another structural constraint that can be used to rule out unrealizable matchings (i.e., point correspondences that cannot be generated by the projection of any 3–D point set). This constraint was introduced by Lu et al. [13] for the particular case of cameras with focal planes parallel to the baseline (and therefore epipoles at infinity and parallel epipolar lines). The idea of [13] is simple: by establishing an order on the epipolar planes, a "natural ordering" of the parallel epipolar lines and thus of the feature points is induced in each image. If the natural ordering pair is known, feature matching becomes a simple alignment process. If not, a candidate matching is rejected because unrealizable if no natural ordering pair exists that aligns the feature points in the two images to create such a matching. Since there are only $\mathcal{O}(N^2)$ possible natural ordering in each image, at most $\mathcal{O}(N^4)$ realizable matchings exist (as opposed to $N!$ possible permutations of the indices).

This paper extend the work in [13] to the general case of epipoles located in finite positions. Two main difficulties arise in this treatment. First, since the epipolar lines in a given image plane form a pencil, natural orderings can only be defined (in general) up to a circular shift, as well as up to order reversal. Second, unlike the case of epipoles at infinity, whereby a natural ordering is uniquely determined by

the direction of traversal of the epipolar lines, in the general case the natural ordering is shown to depend on the *elementary cell* in which the epipole is located. Elementary cells are the faces of a particular arrangement of half–lines in the plane.

1.1. Previous Work

The classic paper by Deriche et al. [5] used a search window for every feature, and a relaxation algorithm to minimize an energy function based on the local structure of the matches. Due to parallax, however, the local feature structure may change considerably. Also, as mentioned earlier, a limited size search window may not work well under wide baseline. A number of papers have appeared that assume "simple" local geometric transformation (affine or homography) [6, 20, 17, 11]. Again, these assumptions may break down in the case of large differences in viewpoint and orientation of the cameras and non–planar scene geometry. A considerable amount of recent work has dealt with the definition and use of invariant and unambiguous features for appearance–based matching [16, 12, 2, 19]. Visual appearance is obviously very relevant for finding correspondences, but suffers from *variation* (change in appearance due to change in geometry or illumination) and *ambiguity* (the fact that different features may look alike). The algorithm proposed here uses visual appearance for the cost function to be minimized, but exploits a structural constraint to alleviate the effect of variation and ambiguity.

When some knowledge about the motion of the cameras is available, it can be used successfully to reduce the search space and improve reliability as in [7]. The combinatorial aspects of point matching have been studied in [14], which formulates the task as an integer optimization problem and proposes a solution based on a the minimization of a concave objective function.

2. Ordering Constraints across Two Views

Consider the two–camera viewing geometry of Fig. 1. The only assumption in our discussion is that the two cameras can be well modeled by a pinhole system. We don't make any assumptions about the location/orientation of the cameras, the geometry of the feature points, and possible occlusions and field of view restrictions. In addition, we don't assume knowledge of the intrinsic parameters of the cameras.

The baseline (the segment joining the two cameras' optical centers) lies on the edge of the pencil of epipolar half–planes, which fully defines the epipolar geometry of the system. The intersection of this pencil of half–planes with each camera's focal plane determines a pencil of epipolar half–lines around each epipole [10].

Suppose that a number of feature points have been identified in the images taken of the same scene by the two cameras. Consider the set of 3–D points that project onto the selected feature points in the two images (each 3–D point projecting onto one point in one or both cameras). These 3–D points can be ordered in the following way. Suppose a "probing" epipolar half–plane, hinging on the baseline, rotates according to an arbitrary (but fixed) direction over the full 360° span. This motion determines the rotation of an epipolar half–line in each image around the corresponding epipole. The direction of rotation (clockwise or counterclockwise) in each image depends on the direction of rotation of the probe and on the orientation of the camera with respect to the baseline.

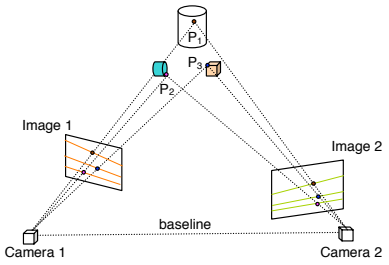

Figure 1. The epipolar geometry of a stereo system.

Suppose for the time being that each feature point in either image has one matching feature point in the other image (we will relax this condition soon). Since the probing half–plane scans the whole 3–D space, it will touch each 3–D point in turn as it rotates. The order in which the points are scanned induces an ordering of their projections in the two focal planes. This simple observation is at the basis of our structured approach to matching. If the location of the epipoles in the two images is known, we can recover the point order in each image (induced by the rotating half–plane) up to order reversal and possibly to a cyclic shift as follows. For each image, consider a rotating half–line hinging on the epipole. This defines an ordering of the feature points, corresponding to the order in which they are touched by the line, which in turn reflects the order in which the corresponding 3–D points are touched by the rotating half–plane. We will call this the "natural ordering" of points induced by the epipole. For example, the rotating half–line shown in Fig. 2(a) determines the following natural ordering: (1,8,2,7,3,5,6,4). Since the direction of rotation of the probing half–line is arbitrary, the ordering is defined up to reversal. In addition, since the starting point is chosen arbitrarily, the ordering of the points in each image is defined up to a cyclic shift. However, it is not always necessary to consider cyclic shifts of the ordered features in the two images. One easily sees that if, in each image, the epipole is outside of the convex hull of the feature points (as in the

case of Fig. 2(a)), then only one feature ordering per rotation direction (clockwise or counterclockwise) of the probing half–line suffices. This is equivalent to the case of parallel epipolar lines considered in [13], whereby only two directions of traversal of the lines needed to be considered.

We thus conclude that the lists of ordered points in the two images are perfectly aligned up to reversal and possibly a cyclic shift. This leads to the following statement:

Proposition 1: Given N feature points in each image of a stereo pair with known epipoles, assume that each point in either image has exactly one (unknown) correspondence in the other image, and that no two points in either image are aligned with the image's epipole. If, for each image, the epipole is outside of the convex hull of the feature points, then there exist at most 2 realizable matchings of the two sets of feature points. Otherwise, there exist at most $2N$ realizable matchings.

If two or more points lie in the same probing half–line, then any permutation of such points generates another potentially realizable matching. By imposing some constraints on the surfaces in the scene, one can reduce the effect of this additional degree of freedom (this is the so–called "ordering constraint" in stereo). In practice, feature points are limited in number, and therefore the likelihood of two or more points aligned with the epipole is negligible. Also, note that the natural ordering *does not* guarantee that all $2N$ candidate matchings are realizable. The correspondences in the matching must also satisfy the epipolar geometry, which may be fixed by as few as seven matches. The natural ordering, however, allows us to rule out a large number of provably unrealizable matchings without the need to estimate the epipolar geometry.

Proposition 1 thus states that, if the epipole positions are known, only 2 or $2N$ candidate matchings of two point sets need to be considered. The determination of the "correct" matching is usually performed based on the visual appearance of the features. We will first consider the case in which the epipole in one or both images lies within the convex hull of the feature points. Let P^i be the ordered point list in the i–th image (with $P^i(n)$ representing the n–th point in the list), and let P^i_k be the cyclic shift of P^i by k elements (where k ranges from 0 to N-1). If $d(p_1, p_2)$ is a positive number representing the "visual dissimilarity" between the descriptors of two points p_1 and p_2, the cumulative dissimilarity can be defined as:

$$C(k) = \sum_{n=1}^{N} d\left(P^1_k(n), P^2(n)\right) \tag{1}$$

The goal is thus to minimize $C(k)$ over k. This operation must be repeated after reversing the order of the points in one of the two sequences, to account for the rotation direction ambiguity.

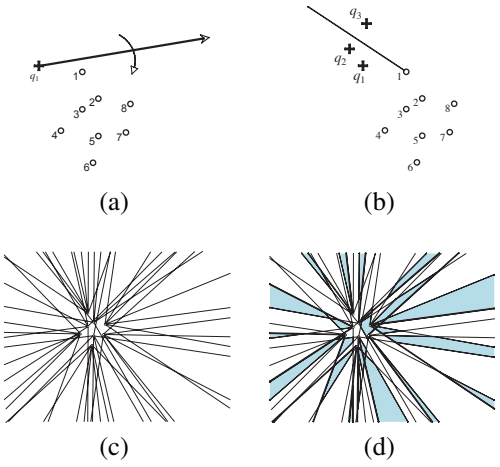

Figure 2. (a): A set of feature points, together with the rotating probing half–line centered on the epipole (marked by a "+"). (b): The critical half–line $l(1, 8)$, together with three possible epipole positions. (c) The arrangement of all critical half–lines (not all elementary cells are visible in the image). (d) A possible subset of elementary cells for coarse–to–fine search.

In general, there will be feature points in one image that don't have a match in the other image (which also implies that the two point sets may have different cardinalities N_1 and N_2). This may be due to limited field of view, occlusions, or imperfect feature detection. In this case, the minimum cost cyclic alignment over all subsequences of P^1 and P^2 should be determined. Let I^i be an ordered subset of indices between 1 and N_i, and denote by $|I^i|$ its cardinality. A common practice is to assign a fixed positive cost c to each deletion of a point from a sequence. Hence, one may seek to minimize the following cumulative cost:

$$C\left(k, I^1, I^2\right) = c \cdot \left(N_1 + N_2 - 2|I^1|\right) + \sum_{n=1}^{|I^1|} d\left(P^1_k(I^1(n)), P^2(I^2(n))\right) \tag{2}$$

over k and all I^1 and I^2 with $|I^1| = |I^2|$.

For a fixed circular shift k, the optimal I^1 and I^2 can be found with complexity of $\mathcal{O}\left(N_1 N_2\right)$ using dynamic programming [4]. Finding the optimal alignment over all possible cyclic shifts using a brute–force approach requires complexity of $\mathcal{O}\left(N_1^2 N_2\right)$ (if $N_1 < N_2$). This complexity is reduced to $\mathcal{O}\left(N_1 N_2 \log N_1\right)$ using the algorithm of [15]. An algorithm with data–dependent complexity that is shown to approach $\mathcal{O}\left(N_1 N_2\right)$ time was presented in [8]. If, for each image, the epipole is outside of the convex hull of feature points, than only two regular dynamic programming tests, each with complexity of $\mathcal{O}\left(N_1 N_2\right)$, are required.

Of course, the assumption that the epipole position is known is unrealistic in practice. Lacking this information, one could explore the focal planes of the two images, probing candidate positions for the epipoles, and finding optimal cyclic alignments for each candidate epipole pair. The algorithm would return the natural orderings of the two point sets that minimize the cost functional in (2), together with candidate locations of the epipoles.

Exhaustive exploration of the focal planes may seem like an overwhelming task. In fact, as we show in the next section, only $\mathcal{O}\left(N^4\right)$ discrete positions need to be considered in each focal plane.

3. Searching for the Epipole Location

Consider again the feature point configuration in a camera's focal plane shown in Fig. 2. A small displacement of the candidate epipole to position q_2 (see Fig. 2(b)) does not affect the natural point ordering. Indeed, it is easy to see that the natural ordering changes only when the candidate epipole crosses a *critical half–line*. A critical half–line $l(p_1, p_2)$ is a half–line departing from p_1, lying on the the line joining p_1 and p_2, and oriented in the opposite direction of p_2. For example, Fig. 2(b) shows the critical half–line $l(1, 8)$. When the candidate epipole crosses $l(p_1, p_2)$, then the points p_1 and p_2 swap position in the natural ordering defined by the epipole. In the case of Fig. 2(b), the new ordering after the epipole crosses $l(1, 8)$ is (8,1,2,7,3,5,6,4). It is also easy to see that crossing a segment joining two feature points does not determine a change in natural ordering (hence the choice of half–lines rather than full lines).

Thus, the *arrangement* of all critical half–lines induced by the feature points determines all the possible natural point orderings. We will call the faces of such an arrangement *elementary cells*. The natural ordering of points does not change as the epipole moves within an elementary cell. Hence, the number of elementary cells represents an upper bound for the number of distinct natural orderings. The arrangement of critical half–lines for the point configuration of Fig. 2(a) is shown in Fig. 2(c).

The main properties of line arrangements in the plane were discovered almost two centuries ago [1]. It is well known that a simple arrangement of M lines (i.e., such that no more than two lines intersect in the same point, and no two lines are parallel) defines $\mathcal{O}(M^2)$ faces, and that an arrangement can be computed in $\mathcal{O}(M^2)$ time. Similar results apply to the case of non–simple half–line arrangements like the one considered here. If there are N feature points in the image, the number of critical half–lines is $N(N-1)$. Hence, there are $\mathcal{O}(N^4)$ elementary cells in a focal plane. Note that two different cells may actually induce the same point ordering. Determining a maximal set of cells such that the induced point orderings are all different requires $\mathcal{O}(N^5)$ time.

Keeping in mind that a set of at most $2N$ candidate matchings is determined by the position of the epipoles in both focal planes (according to Proposition 1), we can state the following:

Proposition 2: Under the same hypotheses of Proposition 1, if the positions of the epipoles are unknown, there exist at most $\mathcal{O}\left(N^9\right)$ realizable matchings of the set of feature points in the two images.

A direct consequence of Proposition 2 is that exhaustive search for the optimal matching requires $\mathcal{O}\left(N^8\right)$ tests (each requiring dynamic programming for determining the optimal alignment) rather than $N!$ Still, this complexity is too high for practical implementations with a reasonable amount of features. In the next section we introduce a coarse–to–fine search strategy to reduce this complexity.

3.1. Coarse–to–Fine Search

The elementary cells in a focal plane can be considered as the nodes of a graph, where two nodes are linked by an edge if the corresponding elementary cells are adjacent. As we saw earlier, moving from one node to a neighboring node determines a 2–point swap in the induced natural ordering. One may expect the change in the minimal cost (2) consequent to a 2–point swap to be small in general. This was shown experimentally for the cases of epipoles at infinity in [13]. This observation suggests the following coarse–to–fine strategy: For each focal plane, select a subset of nodes from the original critical cell graph; find the nodes in the two subsets that minimize the associated cost (2); then, refine the search by exploring nodes that are close to those found in the first step. "Closeness" here is defined in terms of distance in the underlying graph, whereby the distance of two nodes is equal to the length of the shortest path linking them.

Care must be taken to avoid the risk of missing a global minimum, which may happen if the node subset is not dense enough. A sensible criterion for choosing the node subsets is thus to impose a maximum distance k (computed with respect to the original graph) between any given node in the original graph and its closest neighbor in the subset. This distance corresponds to the number of point swaps between the natural orderings induced by the two nodes. In graph theory parlance, we are looking for a *distance–k dominating set* of the original graph [3]. Although optimal algorithms to find distance–k dominating sets are shown to be NP–hard, a number of suboptimal algorithms have been proposed in the literature [3].

The complexity of this coarse–to–fine strategy is proportional to the product of the densities of the node subsets in the two images (where by density we define the

ratio between the cardinality of the subset and the order of the original graph). To the authors' knowledge, there are no published results about the expected density of a distance–k dominating set, although a number of bounds for 2–dominating sets have been reported [18]. We conjecture that the minimum expected density is between $1/2k$ and $1/4k^2$, which correspond to the the minimum densities for a linear chain and for a 8–connected grid respectively.

How large should k be? It seems reasonable to choose a value proportional to the number of feature points: $k = mN$, with $0 < m < 1$ (e.g., m=1/5). Hence, based on our previous argument about the expected density of node subsets, we conjecture that the overall number of candidate epipole location pairs to be tested reduces to somewhere between $\mathcal{O}(N^4)$ and $\mathcal{O}(N^6)$.

4. Experiments

We tested the proposed matching algorithm with a number of image pairs, shown in Fig. 3–5. In all cases the epipoles were in finite position. Given the high algorithm complexity, only a relatively small number of feature points per image was allowed. In the image pairs of Fig. 3–4, points were selected manually. This experiment was meant to validate the algorithm's performances under an ideal feature selection scenario. The left and right images had the same number of points (21 for Fig. 3, 22 for Fig. 4) with one–to–one correspondence. For the pair of Fig. 5, features were automatically chosen using Harris' corner detector [9]. In this case, the left image contained 29 points while the right image contained 30 points. The number of actual correspondence was 24. In all cases, the local appearance of features was represented by Lowe's descriptor [12], computed using a publicly available implementation[1]. The same scale for the descriptor was used for all features in all test images.

Coarse–to–fine search was performed using a very simple (and highly suboptimal) algorithm for selecting the initial subset of cells. A limited number R of critical half–lines (in our experiments, R=16) is chosen, with orientations distributed uniformly in the 2π span. Each selected line is traversed, enumerating the intersections with all the other half–lines of the original arrangement. The elementary cells identified by regular uniform sampling (with density L) of such intersections are used for the initial search. For example, Fig. 2(d) shows a possible subset of elementary cells with L=2 (we used L=4 in out experiments). Once the cells pairs that minimize cost (2) are found, a new set of cells is considered for each image as follows. Given the line l in the subset that bounds the chosen cell, a new subset of half–lines is selected by first finding the two half–lines l_1, l_2 (in the subset of R half–lines) with orientation closest to the orientation of l, and then choosing R new half–lines from the original arrangement of half–lines, such that their orientation is bounded by the orientation of l_1 and l_2. The new optimal cell pair is chosen based on this new subsets in the two images, and the process is repeated iteratively. Future research will investigate the use of more refined algorithms for approximating to the distance–k dominating set as discussed in Sec. 3.1. When candidate cell pairs are outside the convex hull of points within each image, standard dynamic programming is used to find the best alignment and discard outliers. Otherwise, the cyclic–shift dynamic programming algorithm of [8] is used. In our experiments we found that only about one sixth of the cell pairs being tested fell into the second case. Unfortunately, even with this coarse–to–fine implementation, several hours of computation time were required for our experiments.

The matchings found with our algorithm are compared with a simple greedy algorithm that traverses the list of points in one image (ordered arbitrarily) and finds the point in the other image that has the most similar appearance. The number of correct matches found by our algorithm versus the actual number of matches for the cases of Fig. 3–5 is: 19/21; 13/22; 21/24. The greedy algorithm scored: 15/21; 9/22; 18/24. Given the richness of Lowe's local descriptor, the greedy matching algorithm performs fairly well, but our algorithm consistently finds more correct matches.

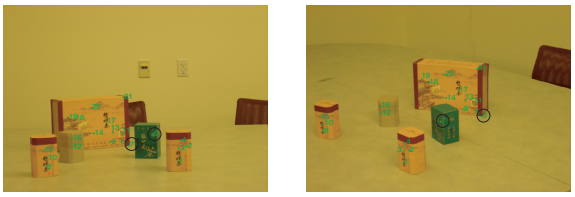

Figure 3. A point matching experiment with the proposed algorithm. The feature points were chosen manually, with local appearance represented by Lowe's descriptor. Both the left and the right image contain 21 points with one–to–one correspondence. Our algorithm found 19 correct correspondences (shown in the figure), while the greedy algorithm found 15 correct ones. Incorrect matches by our algorithm are marked by a black circle.

5. Conclusions

We have introduced a structural constraint that must be satisfied by any realizable matching of image points from two uncalibrated pinhole cameras viewing the same scene. No assumption was made about the geometry of the 3–D points or of surfaces in the scene, nor about the location or orientation of the cameras. This structural constraint can

[1] www.robots.ox.ac.uk/~vgg/research/affine/descriptors.html

Figure 4. A point matching experiment with the proposed algorithm. The feature points were chosen manually, with local appearance represented by Lowe's descriptor. Unmatched points are marked without a number. Both the left and the right image contained 22 points with one–to–one correspondence. Our algorithm found 13 correct correspondences (shown in the figure), while the greedy algorithm found 9 correct ones. Incorrect matches by our algorithm are marked by a black circle, while points that were discarded (unmatched) are marked by a cross.

Figure 5. A point matching experiment with the proposed algorithm. The feature points were chosen by Harris' corner detector, with local appearance represented by Lowe's descriptor. Unmatched points are marked without a number. The left images contained 29 points, while the right image contained 30 points.The number of actual correspondences was 24. Our algorithm found 21 correct correspondences (shown in the figure), while the greedy algorithm found 18 correct ones. Incorrect matches by our algorithm are marked by a black circle, while points that have a correspondence in the other image but were discarded (unmatched) are marked by a cross.

be used to reduce the search space when matching a number of feature points in two images, eliminating the risk of selecting a matching that is provably wrong because unrealizable. We have shown that the maximum number of candidate matchings is $\mathcal{O}(N^9)$, and introduced a coarse–to–fine strategy to reduce this number considerably.

We are well aware that, as it stands, this algorithm is still too computationally heavy for practical use. Yet, we believe that this novel theoretical framework has great potential for future research. For example, it would be interesting to study how any prior knowledge about the relative geometry of the cameras translates into efficiency improvement. We are also considering combining our approach with a RANSAC–type randomization strategy to take epipolar constraints into account.

References

[1] P. Agarwal and M. Sharir. Arrangements and their applications. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*. Elsevier, 2000.

[2] A. Baumberg. Reliable feature matching across widely separated views. In *CVPR00*, pages I: 774–781, 2000.

[3] Y. P. Chen, A. Liestman, and J. Liu. Clustering algorithms for ad hoc wireless networks. In Y. Pan and Y. Xiaoy, editors, *Ad Hoc and Sensor Networks*. Nova Science Publishers, 2004.

[4] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to algorithms*. MIT Press, 1990.

[5] R. Deriche, Z. Zhang, Q. Luong, and O. Faugeras. Robust recovery of the epipolar geometry for an uncalibrated stereo rig. In *ECCV94*, pages A:567–576, 1994.

[6] R. Elias. Wide baseline matching through homographic transformation. In *ICPR04*, pages IV: 130–133, 2004.

[7] T. Goedeme, T. Tuytelaars, and L. Van Gool. Fast wide baseline matching for visual navigation. In *CVPR04*, pages I: 24–29, 2004.

[8] J. Gregor and M. Thomason. Dynamic programming alignment of sequences representing cyclic patterns. *IEEE Trans. PAMI*, 15(2):129–135, Feb. 1993.

[9] C. Harris and . M. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conf.*, pages 147–151, 1988.

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[11] C. Lee, D. Cooper, and D. Keren. Computing correspondence based on regions and invariants without feature extraction and segmentation. In *CVPR93*, pages 655–656, 1993.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004.

[13] X. Lu and R. Manduchi. Wide–baseline feature matching using the cross-epipolar ordering constraint. In *CVPR*, 2004.

[14] J. Maciel and J. Costeira. A global solution to sparse correspondence problems. *IEEE Trans. PAMI*, 25(2):187–199, February 2003.

[15] M. Maes. On a cyclic string–to–string correction problem. *Inform. Processing Lett.*, 35:73–78, 1990.

[16] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, October 2004.

[17] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV98*, pages 754–760, 1998.

[18] N. Sridharan, V. Subramanian, and M. Elias. Bounds on the distance two–domination number of a graph. *Graphs and Combinatorics*, 18:667–675, 2002.

[19] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *IJCV*, 59(1):61–85, August 2004.

[20] J. Xiao and M. Shah. Two-frame wide baseline matching. In *ICCV03*, pages 603–609, 2003.