Wide baseline stereo matching with convex bounded distortion constraints

Meirav Galun*§ Tal Amir*§ Tal Hassner‡ Ronen Basri§ Yaron Lipman§

Abstract

Finding correspondences in wide baseline setups is a challenging problem. Existing approaches have focused largely on developing better feature descriptors for correspondence and on accurate recovery of epipolar line constraints. This paper focuses on the challenging problem of finding correspondences once approximate epipolar constraints are given. We introduce a novel method that integrates a deformation model. Specifically, we formulate the problem as finding the largest number of corresponding points related by a bounded distortion map that obeys the given epipolar constraints. We show that, while the set of bounded distortion maps is not convex, the subset of maps that obey the epipolar line constraints is convex, allowing us to introduce an efficient algorithm for matching. We further utilize a robust cost function for matching and employ majorization-minimization for its optimization. Our experiments indicate that our method finds significantly more accurate maps than existing approaches.

1. Introduction

Finding point correspondences in image pairs of a static scene is a classical problem in stereo and structure from motion (SFM). Finding correspondences in wide baseline setups, i.e., when the cameras' focal centers are distant, is particularly challenging. Images obtained in such setups are generally subject to significant distortion and their content may differ substantially also due to occlusion.

The problem of wide baseline stereo matching has received significant attention in recent years (see a brief review in Section 2). Existing approaches have focused largely on developing better feature descriptors for correspondence and on accurate recovery of epipolar line constraints. However, although challenging, the problem of finding correspondences once the epipolar geometry has been estimated has not yet received sufficient attention.

In this paper we introduce a novel method for finding correspondences in wide baseline image pairs of a static scene. Noting that matching is often ambiguous even when epipolar constraints are taken into account, we propose to address the problem by using deformation maps to model geometric changes along epipolar lines. Specifically, given two images and an estimated fundamental matrix, our algorithm seeks to compute a geometric map that relates the images and satisfies two requirements; First, it should respect the epipolar constraints, and, secondly, we bound the amount of distortion that the mapping can exert locally. We refer to such a map by *epipolar consistent bounded distortion* (EBD) map. Our core theoretical contribution is in showing that, while the set of maps whose distortion is bounded is non-convex, its intersection with maps that satisfy the epipolar constraints (with an ordering assumption [2]) is convex, allowing us to introduce an efficient matching algorithm.

Bounded distortion (BD) maps are continuous, locally injective transformations whose conformal distortion at every point (defined as the condition number of their Jacobian matrices) is bounded. Intuitively, the conformal distortion measures how different the local map is from a similarity transformation, *i.e.*, how much the local aspect ratio is changed. Bounding the conformal distortion is motivated by the following observation. Suppose two cameras are set so that their image planes are parallel (including as special case rectified setups). For any fronto-parallel plane it can be readily verified that its projections onto the two image planes are related by a similarity transformation. Therefore such projections undergo no distortion. Bounding the distortion in these setups therefore limits the slant and tilt of the recovered planes.

To formulate our solution we define a cost function that seeks an EBD map that maximizes the number of matches. We optimize this robust objective using majorization-minimization. The use of a robust objective allows us to recover when certain portions of the images are distorted beyond the bounds allowed by our algorithm or when the set of initial correspondences include outliers. We note that our algorithm both discards outliers from the set of input matches and constructs a dense continuous map that determines the motion of every pixel.

We have tested our method on datasets containing pairs of images with ground truth matches and compared it to several state-of-the-art methods. Our method consistently outperformed these methods.

*Equal contributors
§The Weizmann Institute of Science, Israel
‡The Open University, Israel

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.257

2. Previous work

The problem of wide baseline stereo matching has been approached by a number of studies. Considerable effort has been put into designing better features and descriptors and into utilizing them to estimating the fundamental matrix. Several studies have used affine invariant features [33, 35]. A wide variety of alternatives to the SIFT descriptor [24] have been proposed, emphasizing speed (e.g, the Daisy descriptor [31]) or invariance to extreme transformations such as scale changes [14]. Other studies have utilized line segments [4], regional features (e.g., MSER [15] and texture-based descriptors [28]). [27] groups coplanar points by identifying homographies and uses them to estimate epipolar lines. A few of those descriptors were designed to also account for occlusion (e.g., [31, 32]). Finally, a number of studies have approached the problem from a multiview perspective [29, 9].

Relevant to our work are also generic methods for robust, dense matching, based on a variety of point-feature and regional descriptors, such as the SIFT-flow [23, 22], patchmatch [3], NRDC [13], LDOF [7] and, more recently, SPM [16], as well as models of deformation (e.g., [5, 8, 18]), which can potentially be applied in a wide baseline setting. Another recent study [19] proposed an algorithm for mosaic stitching by finding a map that smoothly departs from a global affine transformation. Our experiments include comparison to [18] and [22] modified to seek matches near corresponding epipolar lines. We show that our method outperforms these techniques, suggesting that our global deformation model is more suitable for wide baseline stereo.

Our deformation model is derived from the work of [20], that proposed an approach for optimizing functionals over bounded distortion mappings using a sequence of convex optimization problems. [21] further used this approach for robust feature matching in general pairs of images (analogous to RANSAC [10], but allowing many degrees of freedom). Our work shows that the set of EBD maps are convex, allowing us to introduce an efficient iterative algorithm.

3. Method

In this section we describe our algorithmic approach to the problem of wide baseline image matching. We assume we are given two images $I, J \subset \mathbb{R}^2$, with their fundamental matrix F either supplied as input or computed automatically, e.g., using RANSAC [10]. Our goal is to find a map Φ from I to J that relates corresponding points in the two images; *i.e.*, for every pair of corresponding points, $(\mathbf{p}, \mathbf{q}) \in I \times J$, the desired map satisfies $\Phi(\mathbf{p}) = \mathbf{q}$. We start with a large set of candidate corresponding pairs of points $(\mathbf{p}_m, \mathbf{q}_m) \in I \times J$, $m = 1, ..., n$ that may contain a significant fraction of outlier matches. Then, we search for a map Φ, from the family of EBD maps, that matches as many pairs $(\mathbf{p}_m, \mathbf{q}_m)$ as possible. Specifically, we aim at optimizing

$$\min_{\Phi} \sum_{m=1}^{n} \|\Phi(\mathbf{p}_m) - \mathbf{q}_m\|_2^0 \tag{1a}$$

$$\text{s.t. } \Phi \in \mathcal{D}_\mu, \tag{1b}$$

where for $\mathbf{v} \in \mathbb{R}^2$ the norm $\|\cdot\|_2^0$ is defined by: $\|\mathbf{v}\|_2^0 = 1$ if $\mathbf{v} \neq \mathbf{0}$, and $\|\mathbf{v}\|_2^0 = 0$ otherwise, and $\mathcal{D}_\mu$ is the set of $\mu-$bounded distortion mappings that respect the epipolar constraints, as defined below. The optimization problem (1) strives to maximize the number of matched pairs under the deformation model, $\mathcal{D}_\mu$. This can be seen by noting that the energy (1a) counts how many pairs $(\mathbf{p}_m, \mathbf{q}_m)$ are not matched by Φ. Similarly to [21], we solve (1) by: 1) computing a set of candidate pairs of correspondences $(\mathbf{p}_m, \mathbf{q}_m)$; and 2) optimizing (1) using an iterative re-weighted least-squares (IRLS) approach. However, differently from previous work, we devise a novel formulation of the bounded distortion deformation model that is shown to be *convex* when matching images under the epipolar constraints. The convex model facilitates the optimization of (1), allows considerably faster optimization times, incorporates epipolar constraints, and does not require convexification. We explain the deformation model next.

3.1. Convex Epipolar BD Deformations

At the core of our method is a convex characterization of the space $\mathcal{D}_\mu$ of EBD deformations. In a nut-shell, $\mathcal{D}_\mu$ is a one parameter family of non-rigid deformations that allow bounded amount (μ) of distortion and respect epipolar constraints. To formulate $\mathcal{D}_\mu$ we introduce a *triangulation* $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$ on image I, where $\mathcal{V} = \{\mathbf{v}_i\} \subset I$ is the vertex set, $\mathcal{E} = \{e_{ij}\}$ the edge set, and $\mathcal{F} = \{f_{ijk}\}$ the triangles (faces).

A mapping $\Phi : I \to J \in \mathcal{D}_\mu$ is represented by prescribing new locations to the vertices of the triangulation in the second image, $\widetilde{\mathcal{V}} = \{\widetilde{\mathbf{v}}_i\} \subset J$. The mapping Φ is defined as the unique piecewise-linear (PL) mapping satisfying $\Phi(\mathbf{v}_i) = \widetilde{\mathbf{v}}_i$. We denote by $\Phi_{ijk} \doteq \Phi|_{f_{ijk}}$ the affine map of the restriction of Φ to the triangle $f_{ijk} \in \mathcal{F}$.

Using the entire collection of PL mappings $\{\Phi\}$ defined on a triangulation $\mathcal{T}$ is way too general as every vertex is allowed to move arbitrarily and in the context of stereo this will allow unreasonable geometries to be considered. Instead, we will restrict our attention to a one parameter family of mapping spaces $\mathcal{D}_\mu$ that translate to a reasonable assumption of the scene's geometry. In particular, in addition to imposing epipolar line constraints, we suggest to bound the deviation of the affine maps Φ_{ijk} from similarity transformations using a parameter $0 < \mu < 1$. We next derive this constraint for a single affine transformation and later show how to set the constraints for the entire triangulation $\mathcal{T}$ to define $\mathcal{D}_\mu$.

3.1.1 Epipolar bounded distortion affine map

We now focus on a single affine map. A general planar affine map can be written uniquely as

$$f(\mathbf{x}) = B\mathbf{x} + C\mathbf{x} + \mathbf{t}, \tag{2}$$

where,

$$B = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}, \quad C = \begin{pmatrix} c & d \\ d & -c \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} t^1 \\ t^2 \end{pmatrix}$$

are a similarity matrix, an anti-similarity matrix (*i.e.*, a reflected similarity), and a translation vector, respectively. Following [20], the determinant of the differential of f, that is $B+C$, takes a diagonal form when expressed in terms of B, C

$$\det(B+C) = a^2 - c^2 + b^2 - d^2 = \frac{\|B\|^2 - \|C\|^2}{2},$$

where $\|\cdot\|$ denotes the Frobenious norm. Hence, the affine map f is orientation preserving if $\|B\| \geq \|C\|$.

The singular values $\Sigma \geq \sigma \geq 0$ of $B + C$ also have a simple form in terms of the matrices B and C

$$\Sigma = \frac{\|B\| + \|C\|}{\sqrt{2}} \quad , \quad \sigma = \frac{|\|B\| - \|C\||}{\sqrt{2}}.$$

The ratio of the maximal to minimal singular values, *i.e.* , Σ/σ, provides a scale invariant measure of deviation from similarity. Restricting f to be orientation preserving and of bounded deviation from similarity can be done by requiring the ratio

$$K_f = \frac{\Sigma}{\sigma} = \frac{\|B\| + \|C\|}{\|B\| - \|C\|}$$

to be non-negative and bounded. Equivalently, we could bound the ratio of the anti-similarity and similarity parts directly, *i.e.*,

$$\mu_f = \frac{\|C\|}{\|B\|} = \sqrt{\frac{c^2 + d^2}{a^2 + b^2}}$$

by

$$\mu_f \leq \mu, \tag{3}$$

where $0 < \mu < 1$. μ is a parameter, $\mu_f = \frac{K_f - 1}{K_f + 1}$, and we name this constraint (3) the μ-bounded distortion constraint. Note that for pure similarity $\mu_f = 0$ and the distortion exerted by the map grows as μ_f is increased. The bounded distortion constraint (3) is not convex and requires some convexification to work with in practice [20]. However, surprisingly, it becomes convex when we intersect this constraint with the epipolar line constraints (assuming epipolar line pairs can be oriented, as we explain below). More generally, when the affine map f is known to map some directed line ℓ_1 (*e.g.*, epipolar line) to another directed line ℓ_2, while preserving the direction, then Eq. (3) can be formulated as a convex constraint in B, C, see Figure 1 for an illustration. We summarize this in the following Proposition.

Proposition 1 *The collection of μ-bounded distortion planar affine transformations that map a directed line ℓ_1 to another directed line ℓ_2 is convex.*

We start by proving the proposition for the case that the directed lines both coincide with the X-axis with the positive direction,

$$\ell_1 = \ell_2 = \ell = \text{span}\{\mathbf{e}_1\}$$

where $\mathbf{e}_1 = (1, 0)^T$. By assumption we have in particular that $f(\mathbf{0}), f(\mathbf{e}_1) \in \ell$ and $\mathbf{e}_1^T f(\mathbf{0}) < \mathbf{e}_1^T f(\mathbf{e}_1)$. This implies that

$$\mathbf{e}_2^T \mathbf{t} = 0\,, \quad d = b\,, \quad a + c > 0 \tag{4}$$

where $\mathbf{e}_2 = (0, 1)^T$. Plugging this into (3), squaring and rearranging we get

$$(1 - \mu^2)b^2 + c^2 \leq \mu^2 a^2. \tag{5}$$

If we show that $a > 0$ then taking the square-root of both sides of (5) leads to a (convex) second-order cone (SOC) constraint,

$$\sqrt{(1 - \mu^2)b^2 + c^2} \leq \mu a. \tag{6}$$

Indeed, since $a + c > 0$ and (5) implies that $|a| > |c|$ we must have $a > 0$. We have therefore shown that any affine map (2) that satisfies the assumption (3) and maps the real axis ℓ to itself by preserving the positive direction has to satisfy (4) and (6). In the other direction, any non-zero affine map that satisfy (4) and (6) maps ℓ to itself while preserving the positive direction (since $a + c > 0$) and satisfies (3).

For general directed lines ℓ_1, ℓ_2 we can represent any affine map f^* satisfying the assumptions of Proposition 1 as

$$f^* = g_2 \circ f \circ g_1^{-1} \tag{7}$$

where g_i, $i = 1, 2$, are similarities that map the X-axis ℓ (with positive direction) to ℓ_i, and f is μ-bounded distortion that maps ℓ to itself while preserving the positive direction as above. Note that this change of coordinates does not change the distortion μ_f of the affine map. Therefore, the collection $\{f^*\}$ of all affine maps satisfying the assumption of the proposition with general lines is convex.

The consequence of this proposition is that the set of μ-bounded distortion affine transformations that map an epipolar line in one image to an epipolar line in another image is convex, provided that the pair of epipolar lines can be oriented. Consider a pair of epipolar lines ℓ_1 and ℓ_2. It can be readily shown that any planar patch in 3D whose front side is visible to both cameras will project to ℓ_1 and ℓ_2 with consistent orientation. We note however that for more general scene structures orientation may not always be preserved. Still, many stereo algorithms assume *ordering* (dating back to [2]). We therefore conclude with the following corollary.

Corollary 1 *The collection of μ-bounded distortion planar affine transformations that map a directed epipolar line ℓ_1 to another directed epipolar line ℓ_2 is convex.*

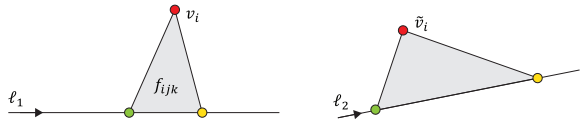

Figure 1. Epipolar bounded distortion affine mapping.

3.1.2 Mappings of triangulations

We use the results of the previous subsection to formulate our convex mapping space $\mathcal{D}_\mu$, where each of its members, $\Phi \in \mathcal{D}_\mu$, is a continuous piecewise linear map whose restriction to any triangle $f_{ijk} \in \mathcal{F}$ is an affine map Φ_{ijk}. Let us denote

$$\Phi_{ijk}(\mathbf{x}) = B_{ijk}\mathbf{x} + C_{ijk}\mathbf{x} + \mathbf{t}_{ijk}.$$

The coefficient of this affine map B_{ijk}, C_{ijk}, and $\mathbf{t}_{ijk}$ are all linear functions of the degrees of freedom $\widetilde{\mathcal{V}}$ (*i.e.*, the mapped vertices) of the mapping Φ as follows,

$$[B_{ijk} + C_{ijk} \mid \mathbf{t}_{ijk}] = \begin{bmatrix} \widetilde{\mathbf{v}}_i & \widetilde{\mathbf{v}}_j & \widetilde{\mathbf{v}}_k \end{bmatrix} \begin{bmatrix} \mathbf{v}_i & \mathbf{v}_j & \mathbf{v}_k \\ 1 & 1 & 1 \end{bmatrix}^{-1} \quad (8)$$

where here $\mathbf{v}_i, \widetilde{\mathbf{v}}_i \in \mathbb{R}^{2\times 1}$ are viewed as vectors in the plane. Note that the inverted matrix (the rightmost matrix in (8)) is constant as it only depends on the source triangulation's vertices $\mathcal{V}$. Therefore, if the triangle f_{ijk} has an edge on an epipolar line ℓ_1, we can set $\ell_2 = F\ell_1$ with F being the *fundamental matrix* and combine (8) with (7), (6) and (4) to constrain Φ_{ijk} to be μ-bounded distortion and to respect the epipolar constraint $\ell_1 \to \ell_2$. See Figure 1 for an illustration. For the third vertex of f_{ijk} (shown in red) we can impose its epipolar constraint by adding the suitable linear equation. Adding these equations for all triangles $t_{ijk} \in \mathcal{F}$ (one SOC and a few linear equality constraints per triangle) results in a convex SOCP realization of the space of PL mappings $\mathcal{D}_\mu$ with a single distortion parameter $\mu \in (0, 1)$.

3.1.3 Triangulating the source image

In order to construct $\mathcal{D}_\mu$ we require a triangulation $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$ with the property that each triangle has an edge on an epipolar line ℓ_1 of image I. We call such a $\mathcal{T}$ an *epipolar triangulation*. We construct such a triangulation by placing an equispaced grid of distance η over a polar coordinate frame centered at the epipole (we used $\eta = 25$ pixels). For each triangle we enforce its edges to coincide with the appropriate epipolar lines by applying constrained Delaunay triangulation. We only keep triangles whose intersection with the image is non-empty. Figure 2 depicts an example. We further determine the orientations of the epipolar lines. This can be done simply by recovering projective camera matrices from the fundamental matrix F and testing the orientation induced, say, by the $Z = const$ plane.

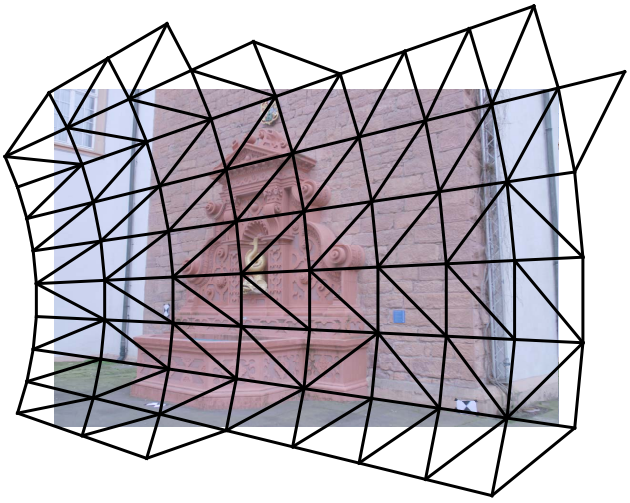

Figure 2. Example of an epipolar triangulation of an image. For illustration purposes we show a coarse triangulation.

3.2. Optimization

To optimize (1) we first use a simple modification of SIFT [24] to find candidate pairs of corresponding points $(\mathbf{p}_m, \mathbf{q}_m)$ that satisfy the epipolar constraint. If the fundamental matrix F is not provided we use standard SIFT and RANSAC to first estimate F.

Next, we optimize (1) using IRLS combined with convex epipolar μ-bounded distortion constraints. Assuming a fixed list of pairs $(\mathbf{p}_m, \mathbf{q}_m)$, we reformulate (1) as

$$\min_{\Phi} \sum_{m=1}^{n} g_{p,\varepsilon}(\|\mathbf{h}_m\|) \quad (9a)$$

$$\text{s.t. } \mathbf{h}_m = \Phi(\mathbf{p}_m) - \mathbf{q}_m\,, \quad m = 1..n \quad (9b)$$

$$\Phi \in \mathcal{D}_\mu, \quad (9c)$$

where $\mathbf{h}_m \in \mathbb{R}^{2\times 1}$ are auxiliary variables, and the functions $g_{p,\varepsilon}$ will be defined soon. The map Φ is represented by the images of the vertices of the triangulation $\mathcal{T}$, that is $\{\widetilde{\mathbf{v}}_i\}$. Namely, each vertex $\mathbf{v}_i$ is mapped to a new (unknown) location in the second image $\widetilde{\mathbf{v}}_i \in J$, and Φ is the unique piecewise linear interpolation Φ_{ijk} over the triangles f_{ijk}, as described in Section 3.1.2. The unknowns in the optimization problem (9) are therefore the target vertex locations $\{\widetilde{\mathbf{v}}_i\}$.

The constraint (9b) is set for every m by finding the triangle f_{ijk} containing $\mathbf{p}_m$ and encoding $\mathbf{p}_m$ in barycentric coordinates of the corners $\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k$ of that triangle, namely $\mathbf{p}_m = c_{m,i}\mathbf{v}_i + c_{m,j}\mathbf{v}_j + c_{m,k}\mathbf{v}_k$, where the barycentric weights satisfy $c_{m,i}, c_{m,j}, c_{m,k} \geq 0$ and $c_{m,i} + c_{m,j} + c_{m,k} = 1$. The image of $\mathbf{p}_m$ under Φ is defined as

$$\Phi(\mathbf{p}_m) = c_{m,i}\widetilde{\mathbf{v}}_i + c_{m,j}\widetilde{\mathbf{v}}_j + c_{m,k}\widetilde{\mathbf{v}}_k. \quad (10)$$

This equation is used in (9b). The EBD constraint (9c) is set by adding Equations (8),(7),(6) and (4) for every triangle

$f_{ijk} \in \mathcal{F}$ of the triangulation $\mathcal{T}$. Note that (6) is a second order cone, and the rest of the equations are linear equalities and inequalities.

Lastly, optimizing the energy (9a) w.r.t. Φ requires to cope with the non-convexity and non-smoothness of the energy (1a). The IRLS point of view suggests replacing the zero norm with its approximations

$$g_{p,\varepsilon}(r) = \begin{cases} r^p & r > \varepsilon \\ \frac{p}{2}\varepsilon^{p-2}r^2 + \left(1 - \frac{p}{2}\right)\varepsilon^p & 0 \le r \le \varepsilon \end{cases} \tag{11}$$

The $g_{p,\varepsilon}$ functions are smooth (C^1) and converge to r^0 as $p, \varepsilon \to 0$. For a fixed p, ε, (9a) is optimized iteratively by replacing $g_{p,\varepsilon}(r)$ with a convex quadratic functional called *majorizer*, $G_{p,\varepsilon}(r, s)$, with the properties that $G_{p,\varepsilon}(s, s) = g_{p,\varepsilon}(s)$, and $G_{p,\varepsilon}(r, s) \ge g_{p,\varepsilon}(r)$, for all r. These two properties guarantee that the IRLS monotonically reduces the energy in each iteration. The majorizers $G_{p,\varepsilon}$ are similar to those in [6],

$$G_{p,\varepsilon}(r, s) = \begin{cases} \frac{p}{2}s^{p-2}r^2 + \left(1 - \frac{p}{2}\right)s^p & s > \varepsilon \\ \frac{p}{2}\varepsilon^{p-2}r^2 + \left(1 - \frac{p}{2}\right)\varepsilon^p & 0 \le s \le \varepsilon \end{cases} \tag{12}$$

Replacing $g_{p,\varepsilon}(\|\mathbf{h}_m\|)$ in (9a) with $G_{p,\varepsilon}(\|\mathbf{h}_m\|, \|\mathbf{h}'_m\|)$, where $\mathbf{h}'_m = \Phi'(\mathbf{p}_m) - \mathbf{q}_m$, and Φ' is the map found at the previous iteration, results in the following convex quadratic energy in $\mathbf{h}_m$ (remember that $\mathbf{h}'_m$ are constants),

$$\min_{\Phi} \sum_{m=1}^{n} w(\|\mathbf{h}'_m\|)\,\|\mathbf{h}_m\|^2 \tag{13a}$$

$$\text{s.t. } \mathbf{h}_m = \Phi(\mathbf{p}_m) - \mathbf{q}_m \tag{13b}$$

$$\Phi \in \mathcal{D}_\mu \tag{13c}$$

where $w(s) = \max\{s, \varepsilon\}^{p-2}$ is constant at each iteration. In view of (10) this implies a convex quadratic energy in the unknowns $\{\tilde{\mathbf{v}}_i\}$. We iteratively solve this problem, updating $\mathbf{h}'_j, \Phi'$ in each iteration until convergence. Each iteration is a convex Second Order Cone Program (SOCP) and is solved using MOSEK [1].

In practice, we fix $p = 0.001$ and ε to be the diameter of image I and solve the above IRLS. Upon convergence, we update $\varepsilon \leftarrow \varepsilon/2$ and repeat. We continue this until $\varepsilon = 1$ (pixels). This heuristic of starting from a large ε and decreasing it helps avoiding local minima of the energy (1a) as the larger the ε the more convex the problem is; for example, for sufficiently large ε the global minimum of (9) lies in the convex (quadratic) part of all terms $g_{p,\varepsilon}$ and can be found by a single SOCP. Our algorithm is summarized in Algorithm 1.

Algorithm 1

Require: Two images I and J, fundamental matrix F, distortion bound μ, edge length η, and a bound on the Sampson distance δ

// Find putative matches
  $\{(\mathbf{p}_m, \mathbf{q}_m)\} = \text{EpipolarSIFT}(I, J, F, \delta)$
// Epipolar triangulation of I according to F
  $\mathcal{T} = \text{DelaunayTri}(I, Constraints(F), \eta)$
Compute barycentric coordinates for $\{\mathbf{p}_m\}_{m=1}^n$
// Optimization
  $p = 0.001, \varepsilon = \text{diameter}(I)$;
$\forall m,\ \mathbf{h}'_m = \mathbf{p}_m - \mathbf{q}_m$
while $\varepsilon \ge 1$ **do**
  while Not converged **do**
    Solve Eq. (13) using SOCP solver, obtaining Φ
    $\forall m,\ \mathbf{h}'_m = \Phi(\mathbf{p}_m) - \mathbf{q}_m$
  end while
  $\varepsilon = \varepsilon/2$
end while
return A subset of matched points $\{(\mathbf{p}_{m_i}, \mathbf{q}_{m_i})\}$ and a map Φ

4. Experiments

Datasets. We evaluate our method by applying the optimization algorithm presented in Sec. 3 to pairs of images from the dataset of [30]. The dataset contains two multi-view collections of high-resolution images (2048×3072), referred to as "Herzjesu" and "Fountain", provided with ground truth depth maps. However, in order to compare to state-of-the-art algorithms, which are considerably slower at those resolutions, we use the lower-resolution (308×461) suggested in [31, 32]. The Herzjesu dataset contains 8 images and the Fountain dataset contains 11 images. Therefore, in total there are 83 stereo pairs with varying distances between focal points. We tested each pair twice, seeking a map from the left image to the right one and vice versa, obtaining 166 matching problems. Note that we do not rectify the images or apply any other pre-processing.

For evaluation we further process the ground truth depth values to obtain ground truth matches. Specifically, for each dataset we employ ray-casting (z-buffering) to the 3D surface, obtaining ground-truth correspondences at sub-pixel accuracy. We further used ray casting to determine an occlusion mask and excluded those pixels (for the left image) from our evaluation. (These masks of course are not known to the algorithm and used only for evaluation.)

Epipolar SIFT. Our algorithm takes as input pairs of putative correspondences and builds an EBD map that is consistent with as many input matches as possible. For the experiments we used SIFT matches (using the VLFeat package [34]). Classical SIFT matching seeks putative matches throughout the *entire* image domain. As we assume that epipolar geometry is known (either exactly or approximately), we modify the matching procedure as follows.

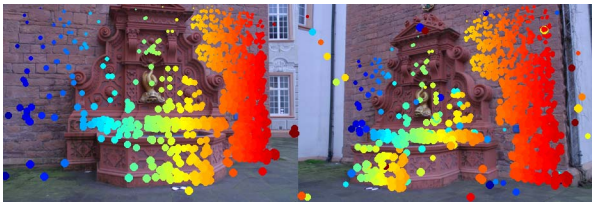

Figure 3. Putative matches obtained with the classical SIFT algorithm, which seeks matches over the entire image. The figure shows images 7 and 3 from the Fountain dataset. Corresponding points are marked by points of the same color and size with color varying with position along the X-axis in the left image and size varying with position along the Y-axis.

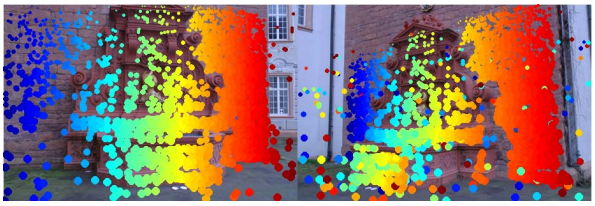

Figure 4. Putative matches obtained with Epipolar SIFT. In this case the search for matches is restricted by the Sampson distance to the immediate surroundings of the corresponding epipolar line. It is evident that the set of putative matches is richer than that obtained with the classical SIFT matching algorithm, Fig. 3.

Figure 5. Matches $\{(\mathbf{p}_m, \mathbf{q}_m)\}$ obtained with our EBD solver.

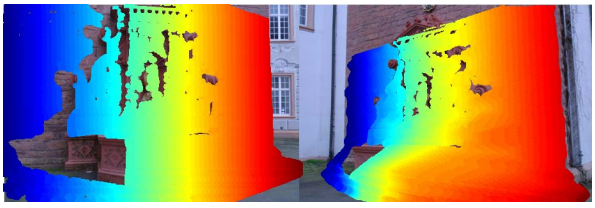

Figure 6. The map Φ obtained with our EBD solver.

Given a SIFT descriptor at location $\mathbf{p}$ in the left image, we restrict the search for a putative match, $\mathbf{q}$, to the area close to the corresponding epipolar line in the right image. This area is determined by limiting the Sampson distance between $\mathbf{p}$ and $\mathbf{q}$, i.e.

$$\frac{(\mathbf{q}^T F \mathbf{p})^2}{(F\mathbf{p})_1^2 + (F\mathbf{p})_2^2 + (F^T\mathbf{q})_1^2 + (F^T\mathbf{q})_2^2} < \delta \qquad (14)$$

where F is the fundamental matrix, $\mathbf{p}$ and $\mathbf{q}$ are written in homogeneous coordinates, and $(F\mathbf{p})_i$ denotes the i^{th} entry of the vector $F\mathbf{p}$. We further accept a match $(\mathbf{p}, \mathbf{q})$ if its SIFT score is at least twice higher (Lowe's criterion) than the score of $(\mathbf{p}, \mathbf{q}')$ for any $\mathbf{q}'$ within Sampson distance δ. We set δ to 5.

The Epipolar SIFT methodology is designed to achieve two objectives. First, it restricts the matches to epipolar lines, and hence removes unnecessary outliers. Secondly, perhaps more importantly, since we only consider matches along epipolar lines each inlier match has fewer competing candidates and so it is more likely to satisfy Lowe's criterion yielding a richer set of putative matches. Fig. 3 shows an example of putative matches obtained using the classical SIFT, while Fig. 4 shows the putative matches obtained with the described methodology, the Epipolar SIFT. It is evident that the set of putative matches obtained with Epipolar SIFT is richer than that obtained with the classical method.

Algorithms for evaluation. We compare our method (EBD) to the following algorithms:

1. **BD**: Feature matching by bounded distortion suggested by Lipman et al. [20]. This method serves as baseline to our method since it seeks correspondences consistent with a bounded distortion transformation, but does not take epipolar constraints into account.

2. **Spectral**: The spectral technique of Leordeanu and Hebert [18]. This method uses graph methods to find point matches by minimizing pairwise energies.

3. **Multi-view Stereo (PMVS2)**: This code, by Furukawa et al. [11, 12], originally designed for multi-view stereo is applied to pairs of images.

4. **SiftFlow**: by Liu et al. [22] finds dense correspondence by minimizing an MRF energy whose unary term measures the match between SIFT descriptors,

5. **Homography**: Mapping by looking for the best homography (computed with RANSAC [10])

6. **Stereo**: An MRF-based stereo algorithm by Lee et al. [17], which finds dense correspondence between the images after rectification.

7. **WxBS-M**: Wide baseline stereo matching by maintaining global (epipolar) and local affine consistency [25]. The method utilizes multiple detector (Hessian-Affine and MSER) and SIFT-based descriptors.

We note that the algorithms of [18] and [22] were not designed specifically for stereo input. For a fair comparison we therefore tested those algorithms in two settings, first in their original (unrestricted) setting, and secondly in a setting that integrates the knowledge of epipolar geometry into the

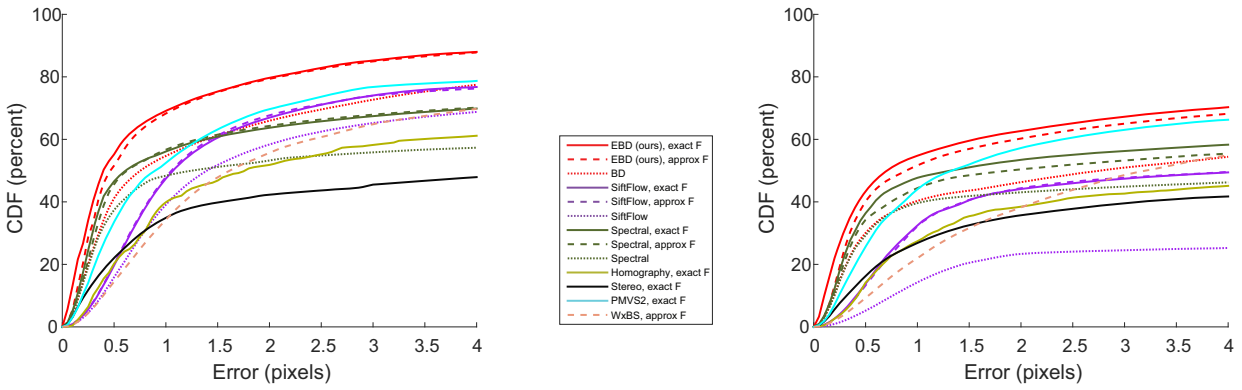

Figure 7. The fraction of pixels mapped by each method to within an error specified on the horizontal axis from their ground truth target location. We present the median for all pairs of images. Dotted lines represent methods disregarding epipolar constraints; dashed lines represent methods using approximate F; solid lines represent methods using exact F, Herzjesu (left) and Fountain (right).

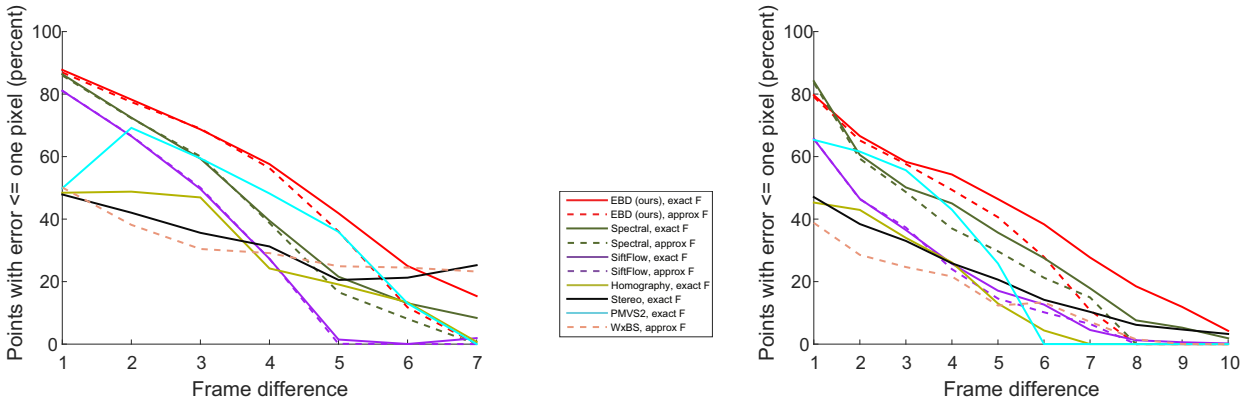

Figure 8. Performance as a function of baseline length. We present the median for all pairs of images, Herzjesu (left) and Fountain (right).

Algorithm	Herzjesu	Fountain
EBD (ours), exact F	**69.11**	**54.77**
EBD (ours), approx F	**68.28**	**51.65**
Spectral, exact F	56.13	47.70
Spectral, approx F	56.70	44.40
PMVS2, exact F	52.47	44.01
SiftFlow, exact F	47.45	32.44
SiftFlow, approx F	47.97	32.19
Homography, exact F	39.95	27.40
Stereo, exact F	34.89	26.84
WxBS-M, approx F	34.30	21.97

Table 1. The fraction of pixels mapped by each method to within one pixel from their ground truth target location. Median computed for all pairs of images in the Herzjesu and Fountain datasets.

algorithms. The latter is achieved as follows. For [18] we used a version of the algorithm that allows it to select from a candidate set of matches that were either extracted from the entire image (for the unrestricted setting) or from the epipolar SIFT matches (*i.e.*, the same input given to our algorithm). Furthermore, since this algorithm does not compute a map (it only return a sparse set of matches) we further applied cubic interpolation to extend the matches to the entire image. (Interpolation was also applied to PMVS2 [11, 12] and WxBS-M [26].) For [22] we modified the code to allow only maps on or close to corresponding epipolar lines (we set the Sampson distance to 2, which gave the best result). For the homography we used putative matches obtained with the epipolar SIFT, and for the stereo algorithm we used ground truth matches to perform the rectification.

Results. Figures 5 and 6 show an example for the results obtained with our method when applied to the input shown in Figure 4. The figures show the set of correspondences $\{\mathbf{p}_m, \mathbf{q}_m)\}$ and the map Φ returned by our optimization, respectively. To further evaluate the map computed with our algorithm for the entire dataset, we checked for each tested pair of images I and J all pixels in I after masking it with the ground truth occlusion map. For each non-occluded pixel $\mathbf{p}$ we measured the Euclidean distance $\|\Phi(\mathbf{p}) - \mathbf{q}\|$,

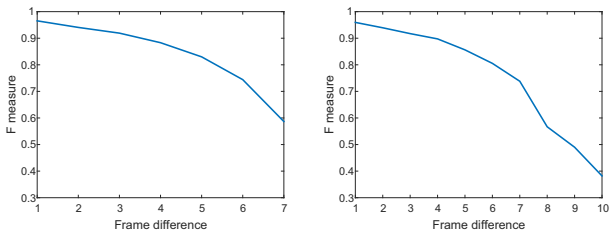

Figure 9. F measure: the success of our algorithm in partitioning the set of candidate matches into inliers and outliers, as a function of the baseline in the sequence for the Herzjesu (left) and Fountain (right) datasets. The inlier ratios are detailed in Table 2.

Frame Diff	1	2	3	4	5	6	7	8	9	10
Herzjesu	0.93	0.89	0.82	0.68	0.52	0.34	0.23			
Fountain	0.96	0.93	0.88	0.82	0.71	0.59	0.43	0.25	0.15	0.085

Table 2. The mean fraction of inliers in the set of putative matches, as a function of the frame difference.

where $\mathbf{q}$ is the ground truth point corresponding to $\mathbf{p}$. We then produced a cumulative histogram depicting the fraction of non-occluded points in I against their displacement error from the ground truth target position. In Figure 7 we report for each error value the median number of points that achieved this error or less over all pairs of images. Table 1 further details the median fraction of non-occluded pixels that were mapped within one pixel accuracy by our map Φ. We show our results both with an exact fundamental matrix (obtained from ground truth) and with an approximated one (computed with RANSAC [10] using classical SIFT). Our results are further compared to Spectral [18], PMVS2 [11, 12], SiftFlow [23] (with and without epipolar constraints), to homography estimation, classical stereo estimation [17] and WxBS-M [26]. (To simplify the table we only include results for the epipolar-enhanced algorithms.) As can be seen from the figures and the table our method outperformed all the tested methods on both datasets with both an exact and an approximate fundamental matrix.

We note further that for all algorithms there was no marked difference between the use of exact and approximate fundamental matrix (solid lines vs. dashed) and all methods benefited from incorporating epipolar constraints (compare to dotted lines, for non restricted version).

Figure 8 further shows a breakdown according to the length of the baseline. For this figure we considered in each of the two datasets all pairs I_i and I_{i+k} for each value k (between 1 and 7 for Herzjesu and between 1 and 10 for Fountain). For each such set of pairs we counted the number of pixels mapped by our computed map Φ with error ≤ 1 pixel and plotted the median of these numbers. As expected the closer together pairs are, the better our method is. Compared to the other methods our method seem to achieve

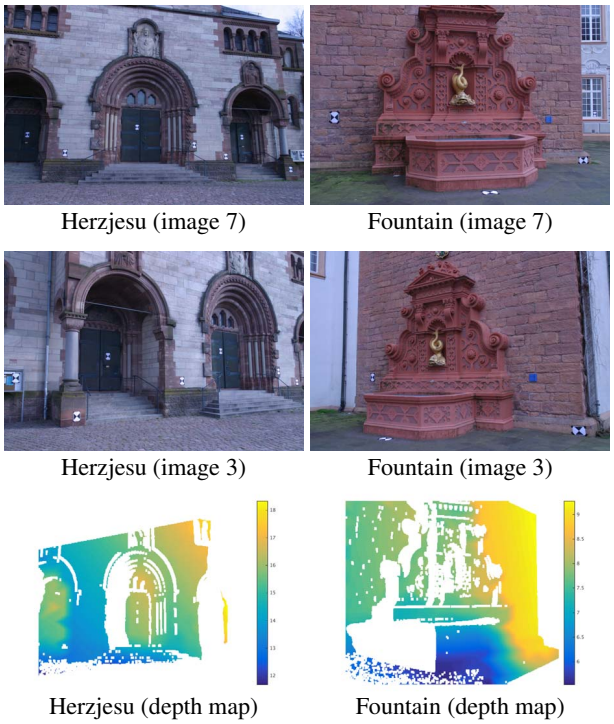

Figure 10. Depth maps computed with our method.

superior accuracy in almost all conditions. We believe that the performance of the method is degraded as the length of the baseline is increased due to the poor number of inlier matches. Table 2 shows the mean fraction of inliers in the set of putative matches as a function of frame difference. Figure 9 shows the success of our algorithm in terms of F measure. The F measure is calculated as follows. The set of candidate matches is partitioned into inliers and outliers (a putative match is considered as an outlier in case its deviation from the ground truth is larger than one pixel). The input to our algorithm is the set of candidate point matches.

Our algorithm filters out inappropriate candidate matches, resulting in a set of inliers and outliers. The accuracy of our partition into inliers and outliers is measured relatively to the partition of the input, yielding the F measure. Finally Figure 10 shows example depth maps computed with our method.

For a pair of images in this dataset our algorithm (non-optimized Matlab code) runs in 70 seconds on a 3.50 GHz Intel Core i7. (The high resolution images require roughly 3.5 minutes.) This is compared to 280 seconds required for the non-convex BD of [20]. In general, running the non-convex BD with features restricted to epipolar lines is significantly slower and achieves slightly inferior results to EBD.

5. Acknowledgements

The research was supported in part by the Israel Science Foundation, Grants No. 1265/14 and 1284/12, I-CORE program of the Israel PBC and ISF (Grant No. 4/11) and the European Research Council (ERC Starting Grant "SurfComp", Grant No. 307754).

References

[1] E. D. Andersen and K. D. Andersen. *The MOSEK interior point optimization for linear programming: an implementation of the homogeneous algorithm*, pages 197–232. Kluwer Academic Publishers, 1999. 5

[2] H. H. Baker and T. Binford. Depth from edge and intensity based stereo. In *Proc. Int. Joint Conf. on Artificial Intelligence*, pages 631–636, 1981. 1, 3

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3), Aug. 2009. 2

[4] H. Bay, V. Ferrari, and L. V. Gool. Wide-baseline stereo matching with line segments. In *CVPR*, 2005. 2

[5] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, pages 26–33, 2005. 2

[6] N. Bissantz, L. Dumbgen, A. Munk, and B. Stratmann. Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces. *SIAM J. on Optimization*, 19(4):1828–1845, 2009. 5

[7] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *CVPR*, pages 41–48, 2009. 2

[8] O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce. A tensor-based algorithm for high-order graph matching. *PAMI*, 33(12):2383–2395, 2011. 2

[9] V. Ferrari, T. Tuytelaars, and L. V. Gool. Wide-baseline multiple-view correspondences. In *CVPR*, 2003. 2

[10] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Com. of the ACM*, 24(6):381–395, 1981. 2, 6, 8

[11] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010. 6, 7, 8

[12] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 6, 7, 8

[13] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Trans. Graph.*, 30(4):70:1–70:9, 2011. 2

[14] T. Hassner, V. Mayzels, and L. Zelnik-Manor. On SIFTs and their scales. In *CVPR*, pages 1522–1528, 2012. 2

[15] M. U. J. Matas, O. Chum and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. 2

[16] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *CVPR*, pages 2307–2314, 2013. 2

[17] S. Lee, J. H. Lee, J. Lim, and I. H. Suh. Robust stereo matching using adaptive random walk with restart algorithm. *Image and Vision Computing*, 37:1–11, 2015. 6, 8

[18] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *ICCV*, volume 2, pages 1482–1489, 2005. 2, 6, 7, 8

[19] W.-Y. Lin, S. Liu, Y. Matsushita, T.-T. Ng, and L.-F. Cheong. Smoothly varying affine stitching. In *CVPR*, pages 345–352, 2011. 2

[20] Y. Lipman. Bounded distortion mapping spaces for triangular meshes. *ACM Trans. Graph.*, 31(4):108:1–108:13, 2012. 2, 3, 6, 8

[21] Y. Lipman, S. Yagev, R. Poranne, D. W. Jacobs, and R. Basri. Feature matching with bounded distortion. *ACM Trans. Graph.*, 33(3):26:1–26:14, 2014. 2

[22] C. Liu, J. Yuen, and A. Torralba. Sift flow: dense correspondence across scenes and its applications. *PAMI*, 33(5):978–994, 2011. 2, 6, 7

[23] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman. SIFT flow: dense correspondence across different scenes. In *ECCV*, pages 28–42, 2008. people.csail.mit.edu/celiu/ECCV2008/. 2, 8

[24] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 4

[25] D. Mishkin, J. Matas, M. Perdoch, and K. Lenc. WxBS: Wide Baseline Stereo Generalizations. In *Proceedings of the British Machine Vision Conference*. BMVA, 2015. 6

[26] D. Mishkin, J. Matas, M. Perdoch, and K. Lenc. Wxbs: Wide baseline stereo generalizations. *CoRR*, abs/1504.06603, 2015. 7, 8

[27] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, 1998. 2

[28] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *ICCV*, 2001. 2

[29] C. Strecha, T. Tuytelaars, and L. V. Gool. Dense matching of multiple wide-baseline views. In *ICCV*, 2003. 2

[30] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, pages 1–8, 2008. 5

[31] E. Tola, V. Lepetit, and P. Fua. Daisy: an efficient dense descriptor applied to wide-baseline stereo. *PAMI*, 32(5):815–830, 2010. 2, 5

[32] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. Moreno-Noguer. Dense segmentation-aware descriptors. In *CVPR*, 2013. 2, 5

[33] T. Tuytelaars and L. J. V. Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC*, 2000. 2

[34] A. Vedaldi and B. Fulkerson. Vlfeat vision software. www.vlfeat.org. 5

[35] J. Xiao and M. Shah. Two-frame wide baseline matching. In *ICCV*, 2003. 2