Robust Guided Matching and Multi-Layer Feature Detection Applied to High Resolution Spherical Images

Christiano Couto Gava, Alain Pagani, Bernd Krolla and Didier Stricker

German Research Center for Artificial Intelligence, Trippstadter Straße 122, Kaiserslautern, Germany
{christiano.gava, alain.pagani, bernd.krolla, didier.stricker}@dfki.de

Abstract: We present a novel, robust guided matching technique. Given a set of calibrated spherical images along with the associated sparse 3D point cloud, our approach consistently finds matches across the images in a multi-layer feature detection framework. New feature matches are used to refine existing 3D points or to add reliable ones to the point cloud, therefore improving scene representation. We use real indoor and outdoor scenarios to validate the robustness of the proposed approach. Moreover, we perform a quantitative evaluation of our technique to demonstrate its effectiveness.

1 INTRODUCTION

The need for generation of accurate 3D models of objects and scenes is increasing as technologies for three-dimensional visualization become more popular and accessible. In this scenario, computer vision algorithms play a fundamental role. Specifically, 3D reconstruction techniques are a promising instrument to support promotion, training, games or education.

Nowadays, image-based reconstruction algorithms are able to produce models of small objects that can compete with those produced by laser scan techniques (Schwartz et al., 2011), (Nöll et al., 2012). These methods demand a highly controlled environment for capturing the images, particularly concerning lighting conditions. Thus they are not suitable for reconstructing scenes in out-of-lab situations.

Nevertheless, reconstruction of large scenes is an attractive tool for documentation, city planning, tourism and preservation of cultural heritage sites (Hiep et al., 2009), (Furukawa et al., 2010), (Pagani et al., 2011). In this context, several reconstruction approaches adopt a region growing strategy, in which 3D points are used as seeds and the scene is gradually reconstructed as regions grow. However, this strategy normally fails when the distance between seed points is large and the final reconstruction is incomplete.

In this paper we present a method that robustly performs matching of image features to support multi-view stereo (MVS) algorithms. Our approach is designed to consistently create seeds and improve scene sampling based on a novel guided matching tech-

nique. It benefits from point clouds produced by modern Structure from Motion (SfM) algorithms and imposes a set of constraints to achieve robustness. Moreover, we propose a multi-layer feature detection method to allow hierarchical matching designed to work with any choice of local feature descriptors.

We apply our algorithm to high resolution spherical images because it has been shown in (Pagani et al., 2011) and (Pagani and Stricker, 2011) that they are more suitable to perform SfM. Due to their wide field of view, these images provide more constraints on camera motion as features are more often observed. Therefore, spherical images are more qualified for guided matching than standard perspective images.

Guided matching has been addressed by other researchers. In (Triggs, 2001) the Joint Feature Distributions (JFD) are introduced. JFD form a general probabilistic framework for muli-view feature matching. The idea is to summarize the observed behaviour of feature correspondences instead of rigidly constrain them to the epipolar geometry. Similar to our work, the method yields confidence search regions instead of searching along the entire epipolar line. In contrast, our approach explicitly combines 3D information with epipolar geometry to define search regions of higher confidence.

The work presented in (Lu and Manduchi, 2004) shares with ours the independence of image features used for matching. Both methods only require a feature detector providing a local descriptor for each feature and a similarity function. However, Lu and Manduchi do not assume calibrated cameras. The method

was designed for the case of nearly parallel epipolar lines, i.e. the epipoles are at infinity. Thus it would face challenging issues with spherical images, because in this case the epipoles are always visible.

The paper is organized as follows: Section 2 introduces the concept of spherical images and related properties. Section 3 outlines our multi-layer feature detection framework. The proposed robust guided matching, our main contribution, is detailed is section 4. Experiments and results are discussed in section 5 and we conclude in section 6.

2 SPHERICAL IMAGES

Spherical images allow to register the entire scene from a single point of view and may be acquired using dedicated hardware and software packages. According to the spherical geometry, each point on the image surface defines a 3D ray r. Analogue to perspective imaging, given a 3D point P_W in world coordinate system (WCS), its counterpart in camera coordinate system (CCS) is $P_C = RP_W + t$, with R and t representing the camera rotation matrix and translation vector. However, different from the perspective case, the relationship between P_C and its projection p onto the image surface is simply $P_C = \lambda p$, with λ being the depth of P_C. Without loss of generality, we assume a unit sphere, leading to $\|p\| = 1$. Thus, in this case, the dehomogenization typical of perspective images is not needed.

Epipolar Geometry: Consider a pair of spherical cameras I_r and I_t. I_r is regarded as reference while I_t is regarded as target camera. Let R_{tr} and t_{tr} be the rotation matrix and the translation vector from I_r to I_t. A point p_r on the surface of I_r, along with the centers of the cameras, define a plane Π, as seen in Figure 1. Π may be expressed by its normal vector $n_\Pi = R_{tr}p_r \times t_{tr}$. For any point p_t on I_t and belonging to Π the condition $n_\Pi^T p_t = 0$ holds. Thus Π is the epipolar plane defined by cameras I_r and I_t. This is the same result obtained in the perspective case and it shows that the epipolar constraint does not depend on the shape of image surface. Nevertheless, to keep consistency with the 3D scene captured by the images, not every point p_t on Π can be a match for p_r. In fact, only those points belonging to the arc defined by p_{t1} and p_{t2} on Figure 1 may be considered for matching. Note that p_{t1} is the epipole on I_t and $p_{t2} = R_{tr}p_r$. We refer to this arc as epipolar arc.

Calibration of Spherical Cameras: Our approach builds on (Pagani et al., 2011) and (Pagani and Stricker, 2011). Thus, we assume the rotation matrix R and the translation vector t for each camera are known. Additionally we assume a set of 3D points resulting from calibration is also provided. This set may be seen as a coarse representation of the scene and is referred to as Sparse Point Cloud (SPC).
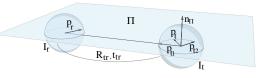

Figure 1: Epipolar geometry of spherical images.

3 MULTI-LAYER FEATURE DETECTION

In this section we focus on the automatic detection of multiple feature layers. The method consists of hierarchically detecting features, thus gradually increasing image sampling. Here the main goal is to support the robust guided matching, which will be detailed in section 4.

Given an image I, a feature detector $\mathcal{F}$ and a parameter vector ρ controlling the behaviour of $\mathcal{F}$, we define a feature layer l as

$$l(I, \mathcal{F}, \rho) = \{f_{u,v} | f_{u,v} = \mathcal{F}(I(u,v), \rho)\}, \quad (1)$$

where $f_{u,v}$ represents a feature detected on image I with pixel coordinates (u, v). To improve readability we will drop the subscripts of $f_{u,v}$ and refer to it as f. $l(I, \mathcal{F}, \rho)$ may be seen as a vector of features, all detected using the parameters ρ. Thus it is possible to define for each image I a set of feature layers L by varying ρ as

$$L(I, \mathcal{F}, \rho_0 \ldots \rho_{K-1}) = \{l_k | l_k = l(I, \mathcal{F}, \rho_k)\}, \quad (2)$$

where $k = 0, 1, .., K - 1$ and K is the number of layers to compute.

Furthermore, we set the parameters $\rho_0 \ldots \rho_{K-1}$ to produce layers with increasing number of features, with the first layer holding the most distinctive features, i.e. the most reliable ones. If, instead of using layers, a single large feature vector is computed, the probability of finding the correct match decreases, because multiple similar features are usually found, i.e. several ambiguous matches are stablished. This is the main motivation to hierarchically create feature layers: They allow dense image sampling without affecting calibration. In other words, with this hierarchical approach, it is possible to:

1. obtain a precise calibration by employing only the first layer(s), i.e. using matches from the most distinctive features;

2. improve performance as less matches need to be computed for calibration;

3. combine as many layers as necessary to perform robust guided matching.

In principle, any feature detector computing the location of the feature on the image along with a local descriptor of its neighborhood could be employed, such as (Lowe, 2004), (Bay et al., 2008) or (Tola et al., 2009). Additionally, a similarity function is required so that descriptors may be compared. In this work, we employ the same feature detector as proposed in (Pagani et al., 2011) and refer to it as Spherical Affine SIFT (SASIFT). SASIFT was chosen due to its robustness against the distortion imposed by the longitude-latitude representation of spherical images. This is specially important near the image poles.

4 ROBUST GUIDED MATCHING

In this section the main contribution of our approach is detailed. The goal is to robustly add 3D points to the SPC to increase the number of seed points for 3D dense reconstruction or to improve the current (sparse) representation of the scene.

Theoretically, an arbitrary number of layers could be computed per image. In practice, few layers are computed because this is already sufficient to achieve both precise calibration – using the first layer – and dense image sampling – using the remaining layers. Yet, these layers may contain several thousands of descriptors and handling numerous images simultaneously is not optimal as computational resources are limited. Thus, we devise the method for pairs of images, so that only the corresponding layers have to be handled. The image pairs are determined according to their neighborhood relation, which is encoded in a binary upper triangular matrix N. If $N(i,j) = 1$, images I_i and I_j are considered as neighbors and matches are computed between them.

Our algorithm combines multiple feature layers, 3D points from calibration and a set of constraints, as epipolar geometry, thresholding and symmetric matching. Moreover, it enforces the consistency of new 3D points and may be applied recursively, allowing to push the number of points even further.

4.1 The Anchor Points

After calibration, most 3D points in the SPC are correctly triangulated. However, some outliers remain. Thus, before applying our guided matching, outliers are removed according to a local density computed for each point in the SPC. We denote the filtered point cloud as S_0. After filtering, all remaining points are assumed to be inliers. These points are regarded as reference and we refer to them as anchor points. We define an anchor point A as a 3D point in Euclidean coordinates along with a set Θ holding the images and the respective features where A is observed.

$$A = \begin{cases} P_W \in \Re^3 \\ \Theta = \{(I_i, f) \mid \lambda p = R_i P_W + t_i\} \end{cases} \quad (3)$$

In Equation 3, p is the image point associated to f. We also define the SPC as the set S of all anchor points. To improve readability we sometimes use A instead of its 3D coordinates P_W throughout the text.

4.2 Matching Based on Anchor Points

In the literature, the term guided matching is usually regarded as the class of methods searching for correspondences given a constraint. This constraint could be imposed by epipolar geometry, a disparity range on aligned images, a predefined or estimated search region or any other criteria that restricts the search for correspondences to a subset of the image pixels.

Our guided matching algorithm is not driven by a single, but by a set of constraints, as described below. Given a reference image I_r, a target image I_t, and a feature f_r detected on I_r, we search for a feature f_t on I_t under the following constraints:

1. Epipolar geometry: $p_r^T E p_t = 0$, with E the essential matrix defined by I_r and I_t, p_r and p_t are the unit vectors corresponding to f_r and f_t;

2. Threshold: the matching score δ between the descriptors of f_r and f_t is above a given threshold τ, i.e. $\delta(f_r, f_t) > \tau$;

3. Symmetry: $\delta(f_r, f_t)$ is the highest score when symmetric matching is performed, that is, f_r and f_t are the best match when the roles of reference and target images are swapped;

However, these constraints are usually not sufficient to achieve robust matching because the set of features f_t complying with the first two criteria above is in general large. As a result, the search has to be done in a large set of potentially ambiguous features.

We propose an approach to overcome this issue. Robustness of guided matching is improved by combining the constraints outlined above, the anchor points and a consistency filter. Our method works as follows: For each feature f_r a set of anchor points projecting on a region Ω centered at p_r is selected. These points form a subset of S, S_Ω. Assuming depth continuity for the points in S_Ω, they can be used to determine a depth range $[\lambda_{min}, \lambda_{max}]$ in which the 3D

point $P_{f_r} = \lambda p_r$ is expected to be. Consequently, the points $P_{min} = \lambda_{min} p_r$ and $P_{max} = \lambda_{max} p_r$ define on the epipolar arc on I_t a confidence region Ψ in which the correct match f_t is expected to be, as shown in Figure 2. This considerably reduces the search region along the epipolar arc, thus increasing the likelihood of finding the correct match. Then we apply the second and third constraints described above to all features in Ψ. Finally, the consistency filter is applied, which will be detailed in section 4.4.
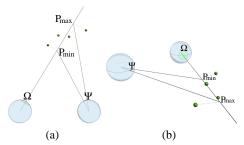

(a)   (b)

Figure 2: Determining the confidence region Ψ based on anchor points. The dots represent the subset S_Ω. (a) Top view and (b) Perspective view.

After applying the consistency filter we consider the following 3 cases: 1. Ψ is empty; 2. Ψ contains one feature; 3. Ψ contains two or more features. In the first case, as no reliable match has been found, the algorithm moves on to match the next feature f_r. For the second case, a new anchor point is created and added to S. Finally, if two or more features remain in Ψ, we discard the matches to enforce robustness, i.e. no anchor point is created.

4.3 Matching Based on Apical Angles

The approach described above works for features f_r whose $S_\Omega \neq \emptyset$. If no anchor points project onto Ω, f_r can not be matched based on anchor points. Increasing the size of Ω does not necessarily solve the issue, as the anchor points in S_Ω may not be representative of the true depth range.

We extend our approach to allow "isolated" features to be matched. Here we do not use anchor points to establish a depth range. Instead, we use apical angles. Apical angle is the angle formed by rays emanating from a 3D point towards the centers of the (pair of) cameras where the 3D point is seen. Then, given a minimum and a maximum apical angle, α_{min} and α_{max}, the depth range is computed as follows. For a feature f_r, we compute the 3D points P_{min} and P_{max} so that the apical angles at P_{min} and P_{max} are α_{max} and α_{min}, as shown in Figure 3. The confidence region Ψ is determined by the projections of P_{min} and P_{max} onto I_t and the rest of the matching proceeds as before.
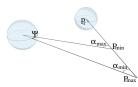

Figure 3: Confidence region Ψ based on apical angles.

4.4 The Consistency Filter

False matches yield 3D points that often violate the ordering assumption. The Consistency Filter supports guided matching by identifying such violations and is explained as follows.

Consider a 3D query point P_q resulting from the triangulation of p_r and p_t. Let A be an anchor point whose projections a_r and a_t onto I_r and I_t are in the vicinity of p_r and p_t, respectively. The vectors p_r and a_r define a normal $n_r = p_r \times a_r$ on I_r. Similarly, $n_t = p_t \times a_t$ on I_t. Given the rotation matrix R_{rt}, P_q is regarded as consistent if $n_r \cdot R_{rt} n_t > 0$ holds for all anchor points projecting in the vicinity of p_r and p_t. Figure 4 shows this concept using a single anchor point.
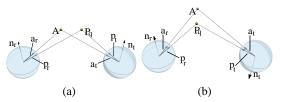

(a)   (b)

Figure 4: Consistency filter. (a) P_q is considered consistent and (b) P_q is considered inconsistent.

5 RESULTS

We applied our algorithm to several different datasets, all captured with images of 100 MegaPixels (approximately 14000 by 7000 pixels). Here we show two of them: the Mogao Cave number 322 in China and the Saint Martin Square in Kaiserslautern, Germany. We processed these datasets on a machine equipped with an Intel® Xeon® CPU W3520 @ 2.67GHz, 24 GB of RAM, running Ubuntu 11.04 - 64 bits.

5.1 Overview

For all datasets we considered the first feature layer for calibration and all layers together for robust matching. As mentioned in section 4.2, our algorithm may be used recursively. We show this by applying the matching based on anchor points in two steps. The first step takes as input the filtered SPC S_0 and outputs a SPC S_1 containing all points in S_0 along with the

new points. Analogously, the second step takes S_1 as input and delivers S_2. We considered two parameters: *first*, a radius r to compute the region Ω centered at the reference feature f_r. We used $r = 100$ pixels in all experiments; *second*, a threshold τ to ensure that only reliable matches are used to add points.

In this work we normalized the SASIFT descriptors. Thus, our similarity function is given by the scalar product between the descriptors of f_r and f_t, i.e. $-1 \leq \delta(f_r, f_t) \leq 1$. The values used for τ were 0.95 and 0.90 in the first and second steps described above, respectively. Values below 0.90 have also been evaluated, but the resulting point clouds started to be corrupted by wrong matches.

Matching based on apical angles takes S_2 as input and delivers a point cloud referred to as S_3. The values used for the angles were $\alpha_{min} = 3°$ and $\alpha_{max} = 45°$. To deal with calibration uncertainties, we considered features located up to 2 pixels away from the epipolar arc during computation of the confidence region Ψ.
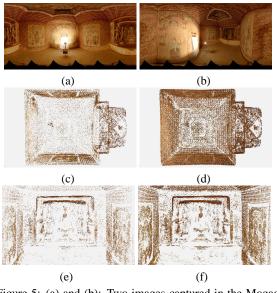

(a) (b)

(c) (d)

(e) (f)

Figure 5: (a) and (b): Two images captured in the Mogao Cave. (c) and (d): Top view of the cave using S_0 and S_2. (e) and (f): Front view of the statues using S_0 and S_2.

5.2 Mogao Cave

This dataset consists of 9 images taken inside the Mogao Cave number 322. In total, 36 pairs were used with 3 layers of features computed for each image. Table 1 summarizes the approximate number of keypoints detected per layer for each image. The third layer contains roughly 5 times the number of features in the first layer, i.e. we considerably increased image sampling. Figure 5 shows the results produced by our algorithm for this dataset. The evolution of

Dataset	L1	L2	L3
Mogao Cave	60000	175000	300000
St. Martin Square	84000	510000	-

Table 1: Approximate number of SASIFT keypoints detected per layer for each image. The lables L1, L2 and L3 identify the corresponding layer.

the number of anchor points n is depicted in Figure 6-(b). The blue curve shows how n raises from S_0 to S_1 during the first matching step. Accordingly, the red curve illustrates the behaviour of n during the second matching step, i.e. from S_1 to S_2.
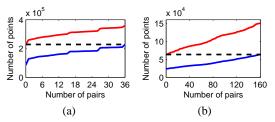

(a) (b)

Figure 6: Number of points versus number of image pairs for the first (blue) and second (red) matching steps. (a) Mogao Cave: S_0, S_1 and S_2 contain 84591, 228044 and 359981 anchor points. (b) Saint Martin Square: S_0, S_1 and S_2 contain 16627, 57706 and 100179 anchor points.

Figure 7 shows the importance of matching based on apical angles. In comparison to Figure 5-(f), the ellipses indicate areas where new points were added. This result was produced by first computing S_3, which delivered 28227 new anchor points, and then applying the matching based on anchor points one more time, with $\tau = 0.90$. The final number of points is 409744.
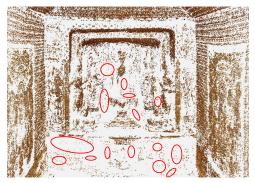

Figure 7: Importance of matching based on apical angles. See text for details.

5.3 Saint Martin Square

The images were taken around a fountain located in the square. This dataset contains 35 images, leading to 161 pairs. Here we computed only 2 layers

of features. Figure 8 illustrates the results regarding the point cloud and Figure 6-(b) shows the evolution of the total number of anchor points for this dataset.
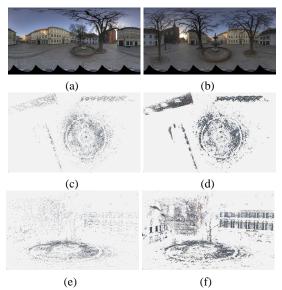

Figure 8: (a) and (b): Two exemplary images taken around the fountain. (c) and (d): Top view of the area using S_0 and S_2. (e) and (f): Close-up on the fountain using S_0 and S_2.

5.4 Reprojection Error

To evaluate the robustness of the proposed approach, we computed and compared the mean reprojection error $\bar{e}$ for S_0, S_2 and S_3. The results are summarized in Table 2. Comparing $\bar{e}(S_0)$ with $\bar{e}(S_2)$, it is clear that our approach reduces $\bar{e}(S_0)$ to roughly $\frac{1}{4}$ of its value. When matching based on apical angles is applied, it further reduces the mean reprojection error for the Mogao Cave dataset. This shows that our approach consistently adds points to the point cloud and improves the initial scene representation.

Dataset	$\bar{e}(S_0)$	$\bar{e}(S_2)$	$\bar{e}(S_3)$
Mogao Cave	7.51	1.90	1.81
St. Martin Square	4.32	0.97	0.97

Table 2: Mean reprojection error $\bar{e}$ computed for S_0, S_2 and S_3, i.e. before and after applying our guided matching technique. Values are given in pixels.

6 CONCLUSIONS

This paper presented a method to robustly add 3D points to sparse point clouds to provide a better representation of the underlying scene. We also proposed a multi-layer feature detection strategy that can be used with several feature detectors and allows features to be hierarchically matched. High resolution spherical images were used as they are more suitable for feature matching. Moreover, our future work includes the development of a dense 3D reconstruction framework based on this type of images.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359.

Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2010). Towards internet-scale multi-view stereo. In *CVPR*.

Hiep, V. H., Keriven, R., Labatut, P., and Pons, J.-P. (2009). Towards high-resolution large-scale multi-view stereo. In *CVPR*, pages 1430–1437. IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60:91–110.

Lu, X. and Manduchi, R. (2004). Wide baseline feature matching using the cross-epipolar ordering constraint. In *CVPR*, volume 1, pages 16–23, Los Alamitos, CA, USA. IEEE Computer Society.

Nöll, T., Köhler, J., Reis, G., and Stricker, D. (2012). High quality and memory efficient representation for image based 3d reconstructions. In *DICTA*, Fremantle, Australia. IEEE Xplore.

Pagani, A., Gava, C., Cui, Y., Krolla, B., Hengen, J.-M., and Stricker, D. (2011). Dense 3d point cloud generation from multiple high-resolution spherical images. In *VAST*, pages 17–24, Prato, Italy.

Pagani, A. and Stricker, D. (2011). Structure from motion using full spherical panoramic cameras. In *OMNIVIS*.

Schwartz, C., Weinmann, M., Ruiters, R., and Klein, R. (2011). Integrated high-quality acquisition of geometry and appearance for cultural heritage. In *VAST*, pages 25–32. Eurographics Association.

Tola, E., Lepetit, V., and Fua, P. (2009). Daisy: An efficient dense descriptor applied to wide baseline stereo. *TPAMI*, 99(1).

Triggs, B. (2001). Joint Feature Distributions for Image Correspondence. In *ICCV*, volume 2, pages 201–208, Vancouver, Canada. IEEE Computer Society.