MULTI-MODAL IMAGE STITCHING WITH NONLINEAR OPTIMIZATION

Arindam Saha, Soumyadip Maity, Brojeshwar Bhowmick

Embedded Systems and Robotics, TCS Research and Innovation, Kolkata, India

ABSTRACT

Despite significant advances in recent years, the problem of image stitching still lacks a robust solution. Most of the feature based image stitching algorithms perform image alignment based on either homography-based transformation or content-preserving warping. Pairwise homography-based approach miserably fails to handle parallax whereas content-preserving warping approach does not preserve the structural property of the images. In this paper, we propose a nonlinear optimization to find out the global homographies using pairwise homography estimates and point correspondences. We further compute local warping based alignment to mitigate the aberration caused by noises in the global homography estimation. To this end, we incorporate geometric as well as photometric constraints to design our cost function which is minimized to obtain better alignment after the global registration, thus producing accurate image stitching. Experimental results on various open datasets demonstrate that our proposed method outperforms state-of-the-art image stitching algorithms.

Index Terms— Image Stitching, Feature, SIFT, Homography, Panorama

1. INTRODUCTION

Image stitching is the procedure of combining multiple images to construct a panoramic image, which is an expression of virtual reality. It is a popular technique to warp a set of visually overlapped images into a single one to obtain a wider field of view (FOV) which can help in many robotic inspection systems like [1, 2, 3]. Image stitching procedure has three major steps - Spatial Calibration, Image Alignment and Blending. The purpose of spatial calibration is to reduce the optical defects and gain correction. The second step calculates the transformation between calibrated image pairs and aligns the images based on the transformation. In the last step, blending technique corrects the misalignment of artefacts. There exists two type of approaches to solve the image stitching problem - direct approach [4, 5, 6, 7] and feature-based approach [8, 9, 10, 11]. The objective of direct approaches is to calculate the suitable homography [12] matrix by minimizing the intensity differences of all overlapped pixels between an image pair. Higher execution time and limited range of convergence [13] are two main disadvantages of direct approaches. Feature-based approaches calculate homography between image pairs using matched features. We chose feature-based approach in this work to stitch images as it is more robust and faster than direct methods. Image alignment using feature-based approach can be categorized in two major ways 1) homography-based transformation [11, 14] and 2) content-preserving warping [15, 16]. The major advantages of homography-based approaches are, warps images globally and thus preserves structural property of images, avoid local distortions. Majority of the homography-based image stitching methods calculate the transformation based on linear algorithms, which ignore few parameters like lens distortion and leads to improper image stitching. But homography based approach miserably fails to handle parallax where the image is non-planar. In this paper, we introduce a novel image stitching method to overcome the above problems. The major contributions of our proposed approach are:

- We propose a nonlinear optimization that estimates accurate homography.
- The proposed multiple constrained local warping method produces robust and better image alignment.

2. RELATED WORKS

Image Stitching is a well-researched topic in the vision community. Very first work on this field is proposed by David L. Milgram [4], where multiple overlapping images are combined into a photomosaic based on geometric registration. Brown *et al.* [11] first propose a fully automatic panorama creation framework, where the authors use object recognition techniques to calculate the overlapping region between two images based on invariant local features. In [14], Brown *et al.* use invariant local features to find matches between images. The authors use the direct linear transformation (DLT) to calculate global homography based on SIFT feature matches. These homography based approaches can't handle large parallax and often generates ghost artefacts. To reduce these ghost artifacts, local warping methods can be useful and Shum *et al.* [15] present a similar work. The authors first align the images based on global homography and then refine that using local warping guided by local motion. Zaragoza *et al.* [17] present a moving DLT based

globally projective warping approach that also allows local non-projective deviations to handle parallax. However, this method is sensitive to the quality and number of feature points. Ziang *et al.* [16] present a hybrid alignment model by combining homographies and content-preserving warping to handle parallax at the same time to avoid objectionable local distortion. The procedure pre-aligns input images using the optimal homography and locally refine the alignment using content-preserving warping. Similarly, Li *et al.* [18] present a warping based motion model incorporating both point and line to preserve geometrical and structural information of scenes. The problem of these approaches is that it strongly depends on the accuracy of feature matching and thus can not handle the noise present in feature correspondences. Photometric constraint is a widely used technique to overcome this scenario. Chen *et al.* [19] present a similar approach combining homography and mesh based local warping. The authors use line and point features to calculate global homography. For refinement, they use a local warping method incorporating photometric as well as geometric constraints. The authors propose a complex and redundant cost function based on point, line, intensity value, mesh, and edge. In this paper, we propose a simpler approach that generates better accuracy.

3. METHODOLOGY

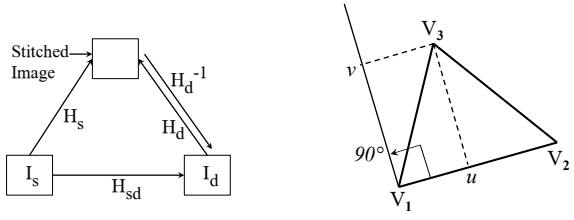
Our proposed approach adopts a dual image alignment method. The first stage estimates a global homography using feature correspondences. This global homography aligns the images and stitches them initially. The second stage estimates local warping using a mesh model in the overlapping regions that smoothen the image boundary for a better stitching. We use pyramidal blending [20, 21] to get final stitched image.

3.1. Global Homography Estimation

We use point features for estimating initial global homography. We use Lowe's SIFT algorithm [22] in order to obtain point correspondences and obtain a set of matched features ψ_1 using the implementation of VLFeat [23]. We also derive another set of matched point correspondence ψ_2 using similar way as explained in [24]. We uses an union of both the point correspondence sets $\varphi_{noisy} = \psi_1 \bigcup \psi_2$ for global homography estimation. A standard practice in the state-of-the-art is to use Direct Linear Transformation (DLT) with random sample consensus (RANSAC) [25] to estimate homography. This type of homography estimation is erroneous and can only stitch images roughly with a lot of misalignments [26, 19]. Therefore, we introduce a further a nonlinear least-square optimization for better estimation of the homography. Let I_s and I_d be an image pair to be stitched and there exists a set θ of such pairs. We estimate all such pairwise homographies using DLT with RANSAC. Each pair generates a pairwise homography matrix H_{sd} and a set of inlier point correspondence φ. Our nonlinear optimization further uses these pairwise homography and inlier point correspondences to estimate global homographies which can stitch a set of images more accurately than pairwise homographies. The cost function is

$$C_\omega = \sum_{I_s, I_d \in \theta} \sum_{i \in \varphi} | \ (H_d^{-1} H_s) x_i - x_i^{'} \ |^2 \ + \lambda * F_R(H_{sd}, (H_d^{-1} H_s)) \tag{1}$$

where x_i and $x_i^{'}$ represent the i-th pair of matching feature point in the images I_s and I_d respectively. H_s and H_d are the global homographies for images I_s and I_d respectively. $F_R(A, B)$ represents the Frobenius norm between matrix A and B. The first part of Equation 1 calculates the error after warping a feature point from the image I_s to the stitched image and subsequently from the stitched image to the image I_d using global homography H_s and H_d. Figure 1(a) explain this warping pictorially where every point correspondence warp from image I_s to image I_d similarly. The second part of Equation 1 constrains the global homography matrices which restricts the homography estimation unboundedly. λ is the balancing weight between these two parts of the Equation 1 and we set to 0.2 in our implementation and give less priority to pairwise homography compared with point correspondences because of erroneous pairwise homography estimation using DLT.



(a) (b)

Fig. 1. (a) Relation between global homographies and pairwise homography matrices of image pair I_s and I_d. (b) V_3 point is linear combination of point V_1 and V_2.

We refer this global homography calculation and alignment as initial alignment where misalignment may exists due to the error in the point correspondences set φ and pairwise homography matrices H_{sd}. The proposed local warping (described in Sec. 3.2) further rectify these misalignments on the overlapping regions as well as the edge boundaries.

3.2. Local Warping

In the local warping step, we concentrate on the overlapped regions between every pair of images. We introduce three different constraints in cost formulation of local warping. Equation 2 represents the cost function for local warping.

$$C_L = C_D + \delta_1 C_P + \delta_2 C_G \quad (2)$$

where C_D, C_P, C_G represent the cost for data term, photometric term, geometric smoothness term respectively. Data and photometric terms tries to rectify the misalignments locally whereas the geometric term ensures a smoothness in object geometry. δ_1 and δ_2 are two balancing weights and we set them to 0.7 and 0.9 respectively in our implementation. We minimize this quadratic cost function and consider to be converged when the average change in pixel movement below to a single pixel.

3.2.1. Data Term

We obtain the feature correspondences between all the aligned image pairs. The locations of these feature correspondences are not exactly same due to the error present in the global alignment. Let an aligned image pair be ${I_s}^{al}$ and ${I_d}^{al}$ which we refer as input images and ${I_s}^{o}$ and ${I_d}^{o}$ are the corresponding output warped images. We take the midpoint x_i^m as the center of i-th point correspondence on input image pair and warp to match with the corresponding midpoint. Equation 3 represents the cost of data term.

$$C_D = \sum_{i \in \varphi} \mid x_i^{os} - x_i^m \mid^2 + \mid x_i^{od} - x_i^m \mid^2 \quad (3)$$

where x_i^{os} and x_i^{od} are the warped feature points in images ${I_s}^{o}$ and ${I_d}^{o}$ respectively.

We take nine corner points $P_{ij}, (j = 1, 2, ..., 9)$, within a 12x12 window around i-th feature correspondence $\forall i \in \varphi$ on input images and represent the i-th feature point with a bicubic interpolation of the nine corner enclosed region. Warped feature points x_i^{os} and x_i^{od} are calculated using Equation 4.

$$x_i^o = \sum_{j=1}^{9} \boldsymbol{w}_{i,j}^T P_{ij}^o \quad (4)$$

where the vector $\boldsymbol{w}_{i,j}$ contains the bicubic interpolation coefficients. P_{ij}^o is the nine corner points on the warped images.

3.2.2. Photometric Term

We introduce a photometric constraint for local warping based on photometric correctness between the image pairs. We create a point set β using sampled points in the overlapped regions and all points on edge boundaries of the overlapping regions. Our cost function matches the intensity of this point in β using Equation 5

$$C_P = \sum_k \mid I_s(x_k^o) - I_d(x_k) \mid^2 \quad (5)$$

where x_k is the k-th point where $k \in \beta$ and x_k^o is the corresponding warped point using a bicubic interpolation. $I_s(.)$ and $I_d(.)$ represents the intensity of source and destination images.

3.2.3. Geometric Smoothness Term

We have added a geometric constraint introducing a unique mesh model. We create a point set ϕ by choosing uniformly sampled points on the edge boundary of the overlapping regions along with matched point correspondences. We further create a triangular mesh by delaunay triangulation [27] using points in ϕ which represents a geometric structure of the overlapping regions. We define the geometric smoothness similar to [28], where any triangle of the mesh is represented as $\triangle V_1 V_2 V_3$ and V_3 is linearly dependent on V_1 and V_2. V_3 can be expressed as

$$V_3 = V_1 + u(V_2 - V_1) + v \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} (V_2 - V_1) \quad (6)$$

where u and v are two scalars can be computed from V_1, V_2 and V_3 as explained in Figure 1(b). The warped triangle $\triangle V_1^o V_2^o V_3^o$ should ensure the similar relation using the u and v in order to achieve a smooth mesh warping. Therefore, the geometric smoothness term is defined as

$$C_G = \sum_{t=1}^{\triangle_n} \mid V_3^o - (V_1^o + u(V_2^o - V_1^o) + v \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} (V_2^o - V_1^o)) \mid^2 \quad (7)$$

where $\triangle_n$ represents the total number of triangle present in the mesh.

4. EXPERIMENTAL RESULTS

We use an Intel i7-8700 (6 cores @3.7-4.7GHz) to implement the proposed approach in C++. The average warp estimation time between a pair of images of around 1024x720 is around 3-4 seconds where majority time is spent on feature detection and matching.

We evaluate the performance of our proposed approach in comparison with three feature based state-of-the-art methods with open dataset used in [17, 18, 19] and captured by ourselves. We present both quantitative and qualitative comparison with previous methods which shows our proposed approach performs better than the state-of-the-art.

4.1. Quantitative Evaluation

We evaluate our proposed approach on publicly available datasets used in [19]. Every data has multiple images with at least 30% overlapping. We exclude low textured data from our evaluation because our proposed approach is feature based. We calculate Root Mean Square Error (RMSE) of one minus normalized cross-correlation (NCC) around a local neighbouring window of 3x3 for all pixels in the overlapping regions between image pairs.

$$RMSE(I_s, I_d) = \sqrt{\frac{1}{\pi} \sum (1 - NCC(x_s, x_d))^2} \quad (8)$$

Data	APAP	DF-W	MCC	Proposed
temple	6.39	3.39	**2.57**	3.105
school	12.20	9.89	10.85	**9.736**
outdoor	11.90	9.52	**6.75**	7.433
rail	14.80	10.58	9.81	**8.317**
building	6.68	4.49	3.74	**3.698**
square	19.90	16.83	12.55	**10.255**
house	19.80	19.57	14.57	**13.113**
courtyard	38.30	36.23	**29.17**	30.258
villa	6.72	**5.20**	5.41	5.332
girl	5.20	4.81	5.05	**4.726**
park	11.07	8.18	**5.85**	7.528
road	2.28	4.59	**1.67**	1.917

Table 1. RMSE error comparison among APAP: as-projective-as-possible method [17]; DF-W: dual-feature method [18], MCC: multiple combined constraint method [19] and our proposed approach.

where π is the total number of pixels in the overlapping regions between images I_s and I_d. x_s and x_d are the pixels in images I_s and I_d respectively.

Table 1 presents the RMSE comparison with previous methods. We can conclude from the comparison that the dual feature method performs better homography estimation than a single point feature. But outdoor datasets where a large amount of point feature presents, feature based homography estimation is very accurate. This indicates the accuracy of global homography estimation increases with the number of matched features. Our proposed approach uses only point features for global homography estimation where feature points are included from both edges and SIFT [22] matching that yields better homography estimation.

4.2. Qualitative Evaluation

Figure 2 shows a qualitative comparison between DF-W [18] and our proposed approach. We execute the comparison on rail data [19] where the main challenge is to merge the rail tracks. Our proposed approach outperforms compare to DF-W [18] in merging the tracks.

Figure 3 presents the qualitative result of our proposed approach which shows global homography estimation using only SIFT [22] point correspondences is erroneous. The edges and overlapping boundaries are not aligned due to lack of matched features in those areas. The result improves significantly by adding more matched point features on edges and overlapping boundaries. Local warping produces the best alignment.

5. CONCLUSION

We present a novel image stitching approach where images are initially aligned using a global homography estimation and further rectify misalignments using a multi-constrained local warping approach. We use photometric as well as geometric constraints in local warping to achieve a smooth structure-preserving stitching among overlapping images. Evaluation of our proposed approach on different open datasets shows better accuracy than state-of-the-art methods.

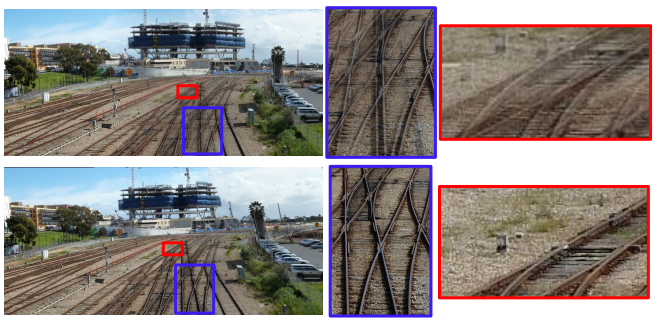


Fig. 2. Qualitative comparison between DF-W [18] and our proposed approach on rail data [19]. First Row: Stitched image using DF-W [18] where misalignment presents in rail tracks. Second Row: Stitched image using our proposed approach where rail tacks are perfectly aligned.

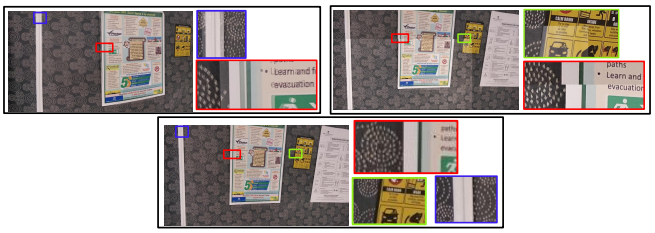


Fig. 3. Qualitative representation of proposed different constraints. First Row Left: Stitched blended image with only global homography using point correspondence set ψ_1 instead of φ_{noisy}. There are highest alignment error; First Row Right: Stitched image with data term and photometric term using point correspondence set φ_{noisy}; The stitched image is without blend to show the misalignment. Second Row: Stitched image with all proposed constraints.

6. REFERENCES

[1] A. Pal, R. Dasgupta, A. Saha, and B. Nandi, "Human-like sensing for robotic remote inspection and analytics," *Wireless Personal Communications*, vol. 88, no. 1, pp. 23–38, 2016.

[2] P. Deshpande, R. V. Reddy, A. Saha, K. Vaiapury, K. Dewangan, and R. Dasgupta, "A next generation mobile robot with multi-mode sense of 3d perception," in *International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 382–387.

[3] A. Saha, S. Maity, and B. Bhowmick, "Indoor dense depth map at drone hovering," in *25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 96–100.

[4] D. L. Milgram, "Computer methods for creating photomosaics," *IEEE Transactions on Computers*, vol. C-24, no. 11, pp. 1113–1119, Nov 1975.

[5] S. Peleg, "Elimination of seams from photomosaics," *Computer Graphics and Image Processing*, vol. 16, no. 1, pp. 90–94, 1981.

[6] D. L. Milgram, "Adaptive techniques for photomosaicking," *IEEE Transactions on Computers*, vol. 26, no. 11, pp. 1175–1180, Nov 1977.

[7] G. Meneghetti, M. Danelljan, M. Felsberg, and K. Nordberg, "Image alignment for panorama stitching in sparsely structured environments," in *Scandinavian Conference on Image Analysis*. 2015, vol. 9127 of *Lecture Notes in Computer Science*, pp. 428–439, Springer.

[8] I. Zoghlami, O. Faugeras, and R. Deriche, "Using geometric corners to build a 2d mosaic from a set of images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[9] P. F. McLauchlan and A. Jaenicke, "Image mosaicing using sequential bundle adjustments," in *Proceedings of the British Machine Vision Conference, BMVC*, 2000, pp. 1–10.

[10] T. J. Cham and R. Cipolla, "A statistical framework for long-range feature matching in uncalibrated image mosaicing," in *IEEE Conference on Computer Vision and Pattern Recognition*. 1998, pp. 442–447, IEEE Computer Society.

[11] M. Brown and D. G. Lowe, "Recognising panoramas," in *IEEE International Conference on Computer Vision*. 2003, vol. 2, p. 1218, IEEE Computer Society.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[13] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, Jan 2006.

[14] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Internatinal Journal on Computer Vision*, vol. 74, no. 1, pp. 59–73, Sep 2007.

[15] H. Y. Shum and R. Szeliski, "Construction and refinement of panoramic mosaics with global and local alignment," in *Sixth International Conference on Computer Vision*. 1998, pp. 953–956, IEEE Computer Society.

[16] F. Zhang and F. Liu, "Parallax-tolerant image stitching," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2014, pp. 3262–3269.

[17] J. Zaragoza, T. Chin, M. S. Brown, and Suter D., "As-projective-as-possible image stitching with moving dlt," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2339–2346.

[18] S. Li, L. Yuan, J. Sun, and L. Quan, "Dual-feature warping-based motion model estimation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4283–4291.

[19] K. Chen, J. Tu, B. Xiang, L. Li, and J. Yao, "Multiple combined constraints for image stitching," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1253–1257.

[20] A. Agarwala, "Efficient gradient-domain compositing using quadtrees," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, pp. 94, 2007.

[21] Z. Farbman, R. Fattal, and D. Lischinski, "Convolution pyramids," in *ACM SIGGRAPH*. 2011, pp. 175:1–175:8, ACM.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[23] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," 2008.

[24] S. Maity, A. Saha, and B. Bhowmick, "Edge slam: Edge points based monocular visual slam," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[25] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun 1981.

[26] C. Herrmann, C. Wang, R.S. Bowen, E. Keyder, M. Krainin, C. Liu, and R. Zabih, "Robust image stitching with multiple registrations," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[27] M. Isenburg, Y. Liu, J. Shewchuk, and J. Snoeyink, "Streaming computation of delaunay triangulations," *ACM Transsaction on Graphics*, vol. 25, no. 3, pp. 1049–1056, Jul 2006.

[28] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, "Content-preserving warps for 3d video stabilization," in *ACM SIGGRAPH*, 2009, pp. 44:1–44:9.