IMAGE ALIGNMENT VIA MULTI-MODEL GEOMETRIC FITTING AND HIERARCHICAL HOMOGRAPHY ESTIMATION

Yue Jiao, Jingyu Yang, Huanjing Yue, Kun Li, Chunping Hou*

Tianjin University, Tianjin 300072, China

ABSTRACT

It is challenging to achieve accurate alignment for building images containing multiple planes. We propose a multi-model geometric fitting and hierarchical homography estimation method to improve the alignment performance for building images. We first extract scale-invariant feature transform (SIFT) features of the images, and then adopt the multi-homography fitting algorithm to classify the feature points into different deformation models. According to the deduced deformation models, we partition the source image into base and transition regions. For the base regions, we adopt the moving direct linear transformation (Moving DLT) to estimate homographies. For the transition regions, we propose a hierarchical homography estimation method to select appropriate homographies. Experimental results show that our method achieves more accurate alignment results compared with state-of-the-art alignment methods for building images.

Index Terms— Image alignment, feature matching, multi-model geometric fitting

1. INTRODUCTION

Image alignment plays a fundamental role in image alignment [15], cloud-based coding [17] and video stabilization [12][2], *etc*. Generally, the alignment performance depends on the accuracy of the estimated homographies. A global homography [15][7] works well for planar scenes or parallax-free camera motions. When the input images violate these imaging assumptions, the global warp may lead to ghosting artifacts or structure distortions. Therefore, the local warping models [11][18] have been emerging in recent years.

Many local warping models have been proposed in recent years. Lin *et al*. proposed a smoothly varying affine (SVA) warping method to handle local deformations while still preserving global affinity [11]. Due to the insufficient degrees of freedom of the affine model, the SVA cannot achieve projective warping. An extension of SVA is proposed by Zaragoza *et al.*, called as-projective-as-possible (APAP), which achieved a smoothly varying projective stitching by moving direct linear transformation (DLT) [18].

Fig. 1: Illustration of the base (B) and transition (T) regions.

Hereafter, Chang *et al*. presented a shape-preserving half-projective (SPHP) warping to achieve gradually change from projective to similarity [4]. Though it can reduce the distortions and preserve the image shape, it is sensitive to parameter selection. To suppress the perspective effect, Lin *et al*. proposed an adaptive as-natural-as-possible (ANAP) warping by linearizing the homography in the nonoverlapping regions while combining these homographies with global similarity transformation [9]. Though it is robust to parameter selection, there are some local distortions in stitched images. Recently, Xiang *et al*. proposed a local warping combing line constraints into a global similarity transform, which could keep the content-consistence of the image and mitigates projective distortions [16]. Lin *et al*. proposed a seam-guided local alignment method for large parallax image stitching [10]. This approach mainly coupled the local alignment computation and used the seam estimation via adaptive feature weighting. Li *et al*. proposed a quasi-homography warp to balance perspective distortion against projective distortion in the nonoverlapping region [8]. Shi *et al.* presented a multi-model method to improve the alignment performance for image set compression [14]. We can see that there's a trend to develop more delicate warping functions for better alignment results.

Since the alignment for images composed by multiple planes cannot be modeled well with a global homography or multiple equally divided local homography transforms, in this paper we propose a multi-model geometric fitting and hierarchical homography estimation method to improve the alignment performance.

*Corresponding author: dayueer@tju.edu.cn. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771339, Grant 61672378, Grant 61571322 and Grant 61520106002.

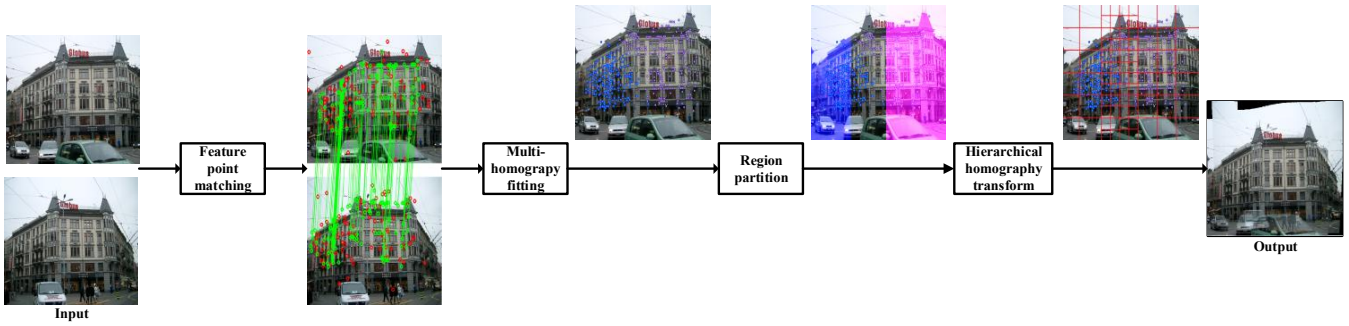

Fig. 2: Framework of the proposed method.

2. THE PROPOSED METHOD

The framework of our method is illustrated in Fig.2. First, SIFT features are extracted and matched between the source and target images. Then, these matching points are divided into multiple groups utilizing the multi-homography fitting method. Then, the source image is partitioned into base and transition regions according to the matched feature groups. Finally, the homographies for the base regions are estimated with the moving DLT and the homographies for the transition regions are calculated by the proposed hierarchical homography transform. These modules are detailed in the following subsections.

2.1. Multi-homography Fitting and Region Partition

2.1.1. Multi-homography Fitting

We first extract their SIFT features and their matching feature points from the source image I and the target image I' using the matching criteria proposed in [13]. Then we use the RANSAC algorithm [6] to remove the outliers. We denote a set of the matched feature points in source and target images as $\Omega = \{\mathbf{x}_i\}$ and $\Omega' = \{\mathbf{x}'_i\}$, respectively. The geometric relationships between different feature points in Ω are modeled via triangulation mesh. We adopt the graph-cut algorithm [3] to classify these matched points into different groups. The objective function is defined as

$$E(\mathcal{H}) = E_D(\mathcal{H}) + \lambda E_S(\mathcal{H}), \tag{1}$$

where, $\mathcal{H} = \{\mathbf{H}_i\}$, $E_D(\mathcal{H})$ is the data term to minimize the distance between transformed $\mathbf{x}$ and $\mathbf{x}'$. $E_S(\mathcal{H})$ is the smooth term, which constraints neighboring feature points have the same homography. λ is the weighting parameter to balance the weight between data and smooth terms. It is set to 10 in this paper. The data term is defined as

$$\begin{aligned} E_D(\mathcal{H}) &= \sum_i \mathcal{D}(\mathbf{H}_i, \mathbf{x}_i, \mathbf{x}'_i), \\ \mathcal{D}(\mathbf{H}_i, \mathbf{x}_i, \mathbf{x}'_i) &= \|\Phi(\mathbf{H}_i\tilde{\mathbf{x}}_i) - \mathbf{x}'_i\|_2^2, \end{aligned} \tag{2}$$

where $\tilde{\mathbf{x}}_i = [\mathbf{x}_i^\top, 1]^\top$ is $\mathbf{x}$ in homogeneous coordinates. $\Phi(\cdot)$ represents converting the homogeneous coordinates to inhomogeneous coordinates. $\mathcal{D}(\mathbf{H}_i, \mathbf{x}_i, \mathbf{x}'_i)$ measures the distance between the warping point of $\mathbf{x}_i$ and its matching point $\mathbf{x}'_i$. $\mathbf{H}_i$ is a homography with size 3×3. The smooth term is defined as

$$E_S(\mathcal{H}) = \sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} \mathcal{S}(\mathbf{H}_i, \mathbf{H}_j), \tag{3}$$

where $\mathcal{N}(\mathbf{x}_i)$ is the neighboring feature points of $\mathbf{x}_i$ and the neighboring relationship is defined by the triangulation mesh. The penalty function $\mathcal{S}(\mathbf{H}_i, \mathbf{H}_j)$ equals 1 if $\mathbf{H}_i \neq \mathbf{H}_j$, otherwise $\mathcal{S}(\mathbf{H}_i, \mathbf{H}_j) = 0$.

The minimization of the energy function in Eq. (1) can be approximated by the graph-cut algorithm. The data node in the graph is the feature point $\mathbf{x}_i$ and the label for each node is the homography $\mathbf{H}_i$. The edge weights between data nodes and labels are defined by the data term in Eq. (2). The edge weights between data nodes are defined by the smooth term defined in Eq. (3). The label candidates $\mathbf{H}_i$ is generated by randomly selecting four matching pairs to calculate one homography. In this paper, we initialize 3000 homographies as labels.

As a result, the feature points of the two images are classified into different groups using the above method. As shown in Fig.1, the feature points in the source image are classified into two different groups, denoted by purple and blue points respectively.

2.1.2. Region Partition

After multi-homography fitting, each feature point in the source image is assigned a unique label, namely, that the feature points in Ω are classified into several regions. Suppose there are M regions, and we denote the *i-th* feature point in the *m-th* region as $\mathbf{x}_{m,i}$. Then the boundary for each region is calculated as

$$\begin{aligned} &\begin{cases} B'_{m,l} = \max\{\min\{\mathbf{x}_{m,i} - 3s(\mathbf{x}_{m,i})\}, 1\}, \\ B'_{m,r} = \min\{\max\{\mathbf{x}_{m,i} + 3s(\mathbf{x}_{m,i})\}, W\}, \\ \mathbf{x}_{m,i} \in \Omega_m, \end{cases} \\ &\begin{cases} B_{m,r} = \min\left\{B'_{m,r}, B'_{m+1,l}\right\}, \\ B_{m,l} = \max\left\{B'_{m,l}, B'_{m-1,r}\right\}, \end{cases} \end{aligned} \tag{4}$$

where Ω_m is the feature point set in the *m-th* region and $s(\mathbf{x}_{m,i})$ is the scale of the feature point $\mathbf{x}_{m,i}$. W is the width

of the source image. The multiplication factor before $s(\mathbf{x}_{m,i})$ is set as 3 because the SIFT descriptor is calculated in a region with radius that equals to $3s(\mathbf{x}_{m,i})$. $B_{m,l}$ is the left (top) boundary, and $B_{m,r}$ is the right (bottom) boundary for the *m-th* region. These regions are named as base regions, and the regions between neighboring base regions are named as transition regions. Two region partition examples are presented in Fig. 1. Considering that the building images are usually captured with the sky at the top and the ground at the bottom of the image, the boundaries in this work mainly means the left and right boundaries.

2.2. Hierarchial Homography Estimation

2.2.1. Homography Estimation in Base Regions

The base region is still not a perfect planar scene. Therefore, utilizing a basic homography to warp the base region will produce misalignment. Therefore, we adopt the moving DLT algorithm [18], which estimates location dependent homography, to warp the base region. Each base region is divided into $C_1 \times C_2$ cells. A point $\mathbf{x}_*$ in the source image I is warped to the position $\hat{\mathbf{x}}_*$ in the target image I' by

$$\hat{\mathbf{x}}_* \sim \mathbf{H}_* \tilde{\mathbf{x}}_*, \tag{5}$$

where $\tilde{\mathbf{x}}_*$ is the homogeneous coordinates of $\mathbf{x}_*$. The homography $\mathbf{H}_*$ for current cell is obtained by minimizing

$$h_* = \arg\min_h \sum_i \|w_*^i a_i h\|^2 \quad \text{s.t. } \|h\| = 1, \tag{6}$$

where h_*, of size 9×1, is the vectorized version of $\mathbf{H}_*$. a_i is derived by the formula $\hat{\mathbf{x}}_i \times \mathbf{H}_i \tilde{\mathbf{x}}_i = \mathbf{0}_{3\times 1}$. The weight parameter w_*^i is defined as

$$w_*^i = \max(\exp(-\|\mathbf{x}_* - \mathbf{x}_i\|^2/\delta^2), \gamma), \tag{7}$$

where $\mathbf{x}_i$ is the *i-th* feature point in current region and $\mathbf{x}_*$ is the center point of current cell. δ is a scaling parameter, and γ is introduced to prevent numerical problems (e.g. poor data). In this way, the homography in each cell is locally adapted to its content and the homographies in neighboring cells are varying smoothly. For more information, please refer to [18].

2.2.2. Homography Estimation in Transition Regions

The transition regions may contain multiple planes, which need more delicate homographies to warp them. We first utilize moving DLT to generate homography candidates for the transition region and then utilize graph cut to choose the most suitable homography for each cell, resulting in hierarchical homographies.

For the transition region T_m between the *m-th* and *(m+1)-th* base regions, we denote its candidate matching feature points as $\Omega_{T_m} = \Omega_m \bigcup \Omega_{m+1}$. We gradually select feature points from Ω_{T_m} and utilize these feature points to generate candidate homographies via moving DLT. We denote the candidate homographies for transition region T_m as $\mathcal{H}_{T_m}$. Hereafter, we divide the transition region T_m into $C_{t_1} \times C_{t_2}$ cells. For each cell, we choose its most suitable homography from $\mathcal{H}_{T_m}$ via graph cut. The objective function is defined as

$$E(\mathcal{H}_{T_m}) = E_D(\mathcal{H}_{T_m}) + \beta E_S(\mathcal{H}_{T_m}). \tag{8}$$

The data term $E_D(\mathcal{H}_{T_m})$ is defined as

$$E_D(\mathcal{H}_{T_m}) = \sum_k \sum_j \|\eta w_{j,k} \mathcal{D}(\mathbf{H}_k, \mathbf{x}_j, \mathbf{x}'_j)\|_2^2, \tag{9}$$

where $\mathbf{H}_k \in \mathcal{H}_{T_m}$ represents the homography for the *k-th* cell and $\mathbf{x}_j \in \Omega_{T_m}$. η is a constant parameter, which is set to 100 in our experiments. The distance function $\mathcal{D}(\mathbf{H}_k, \mathbf{x}_j, \mathbf{x}'_j) = \|\Phi(\mathbf{H}_k \tilde{\mathbf{x}}_j) - \mathbf{x}'_j\|_2^2$ is the same as that defined in Eq. (2). The weighting parameter $w_{j,k}$ is introduced to adjust the penalty of the feature points warping accuracy according to their distance to the *k-th* cell. It is defined as Eq. (10),

$$\begin{cases} \tilde{w}_{j,k} = U_k - \|\mathbf{p}_k - \mathbf{x}_j\|^2 + 1, \\ U_k = \max\{\|\mathbf{p}_k - \mathbf{x}_j\|^2\}, \\ w_{j,k} = \dfrac{\tilde{w}_{j,k}}{\sum_j \tilde{w}_{j,k}} \end{cases} \tag{10}$$

where $\mathbf{p}_k$ is the center point of the *k-th* cell.

The smooth term is defined as

$$E_S(\mathcal{H}_{T_m}) = \sum_{t \in \mathcal{N}(k)} \|\mathbf{H}_t - \mathbf{H}_k\|_1, \tag{11}$$

where $\mathbf{H}_t$ is the homography for the neighboring cell of the *k-th* cell. The smooth term weighting parameter β is set to 10 for boundary cells, otherwise β is set to 0.01. Utilizing the graph cut algorithm to solve Eq. (8), we obtain the homography for each cell. Note that, here we utilize l_1 norm constraint instead of the penalty function utilized in Eq. (3) because the homography matrixes in transition regions should be smoothly varied other than have hard discontinuities.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed method by comparing with APAP [18], SPHP [4], the alignment method in video stablization (AMVS) [12], and GSP [5], since the four methods have released codes. We utilize the images in Zurich building database as our test set [1]. Our method takes about 40 to 50 seconds to align two images with resolution of 640×480 using Matlab on a PC with Intel i5 3.0 GHz CPU and 6 GB RAM.

Fig. 3 presents the alignment results in terms of the superposition of the warped source image and target image for three test images. It can be observed that the compared four methods have ghosting artifacts to some extent since their transformations are not finetuned according to image planes. For example, for the first image, SPHP and AMVS generate obvious ghosting artifacts for the letters on the building. For the second image, all the four compared methods have ghosting artifacts on the letters in the base region, while our method

Fig. 3: Comparison of image alignment results in terms of the the superposition of the warped source image and target image. From left to right, the results are generated by APAP, SPHP, AMVS, GSP, and the proposed method. To show the details clearly, we zoom in the crops highlighted by color boxes for each image.

produces accurate alignment result in this region, since our homographies in base regions are calculated without the disturbance from the feature points in other regions. For the third image, the first three compared methods have ghosting artifacts in the transition region. Thanks to our hierarchical homography transform estimation, our method generates accurate alignment result in transition regions, comparable with that of GSP.

4. CONCLUSION

In this paper, we propose a novel image alignment method for building images containing multiple planes, which cannot be aligned well using single homography. We first partition the source image into base and transition regions according to the matched feature groups, which are derived via graph-cut algorithms. Then, for the base regions, we calculate the transformations via moving DLT algorithm. For the transition regions, we propose a hierarchical homography estimation method to refine the initialized homographies. Experiment results demonstrate that our method achieves state-of-the-art alignment performance for building images. Our work also has limitations. The proposed method may fail to detect small base regions, which have few matching points.

5. REFERENCES

[1] Zurich building image database. [online]. http://www.vision.ee.ethz.ch/showroom/zubud/.

[2] C. Bampis, G. Somanath, O. Nestares, and J. Yao. Panoramic background estimation from rgb-d videos. *Electronic Imaging*, 2017(15):14–19, 2017.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[4] C. H. Chang, Y. Sato, and Y. Y. Chuang. Shape-preserving half-projective warps for image stitching. In *Computer Vision and Pattern Recognition*, pages 3254–3261, 2014.

[5] Y.-S. Chen and Y.-Y. Chuang. Natural image stitching with the global similarity prior. In *European Conference on Computer Vision*, pages 186–201. Springer, 2016.

[6] M. A. Fischler and R. C. Bolles. *Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography*. ACM, 1981.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[8] N. Li, Y. Xu, and C. Wang. Quasi-homography warps in image stitching. 2017.

[9] C. C. Lin, S. U. Pankanti, K. N. Ramamurthy, and A. Y. Aravkin. Adaptive as-natural-as-possible image stitching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1163, 2015.

[10] K. Lin, N. Jiang, L. F. Cheong, M. Do, and J. Lu. *SEAGULL: Seam-Guided Local Alignment for Parallax-Tolerant Image Stitching*. Springer International Publishing, 2016.

[11] W. Y. Lin, S. Liu, Y. Matsushita, T. T. Ng, and L. F. Cheong. Smoothly varying affine stitching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 345–352, 2011.

[12] S. Liu, L. Yuan, P. Tan, and J. Sun. Bundled camera paths for video stabilization. *ACM Transactions on Graphics (TOG)*, 32(4):78, 2013.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[14] Z. Shi, X. Sun, and F. Wu. Multi-model prediction for image set compression. In *Visual Communications and Image Processing*, pages 1–6, 2014.

[15] R. Szeliski. *Image Alignment and Stitching*. Springer US, 2006.

[16] T. Xiang, G.-S. Xia, L. Zhang, and N. Huang. Locally warping-based image stitching by imposing line constraints. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 4178–4183. IEEE, 2016.

[17] H. Yue, X. Sun, J. Yang, F. Wu, et al. Cloud-based image coding for mobile devices-toward thousands to one compression. *IEEE Transactions on Multimedia*, 15(4):845–857, 2013.

[18] J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter. As-projective-as-possible image stitching with moving dlt. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2339–2346. IEEE, 2013.