3D Res (2017) 8:24
DOI 10.1007/s13319-017-0134-y

3DR EXPRESS

Data-Driven Neural Network Model for Robust Reconstruction of Automobile Casting

Jinhua Lin · Yanjie Wang · Xin Li · Lu Wang

Received: 24 April 2017 / Revised: 12 June 2017 / Accepted: 15 June 2017 / Published online: 23 June 2017

Abstract In computer vision system, it is a challenging task to robustly reconstruct complex 3D geometries of automobile castings. However, 3D scanning data is usually interfered by noises, the scanning resolution is low, these effects normally lead to incomplete matching and drift phenomenon. In order to solve these problems, a data-driven local geometric learning model is proposed to achieve robust reconstruction of automobile casting. In order to relieve the interference of sensor noise and to be compatible with incomplete scanning data, a 3D convolution neural network is established to match the local geometric features of automobile casting. The proposed neural network combines the geometric feature representation with the correlation metric function to robustly match the local correspondence. We use the truncated distance field(TDF) around the key point to represent the 3D surface of casting geometry, so that the model can be directly embedded into the 3D space to learn the geometric feature representation; Finally, the training labels is automatically generated for depth learning based on the existing RGB-D reconstruction algorithm, which accesses to the same global key matching descriptor. The experimental results show that the matching accuracy of our network is 92.2% for automobile castings, the closed loop rate is about 74.0% when the matching tolerance threshold τ is 0.2. The matching descriptors performed well and retained 81.6% matching accuracy at 95% closed loop. For the sparse geometric castings with initial matching failure, the 3D matching object can be reconstructed robustly by training the key descriptors. Our method performs 3D reconstruction robustly for complex automobile castings.

Keywords 3D reconstruction · 3D scanning · Data-driven · Learning model · Neural network · RGB-D

1 Introduction

In the field of computer vision and graphics, it is important to study geometric matching for many aspects including 3D reconstruction, object location and tracking. Since the matching performance of local fusion fragment directly affects the reconstruction precision, most advanced 3D reconstruction methods are focused on the improvement of the geometric matching algorithm. However, due to the low-resolution characteristics of the depth data, it is necessary to

J. Lin (✉) · X. Li · L. Wang
Computer Application Technology, Changchun University of Technology, No. 229, Xiuzheng Road, Changchun City 130012, Jilin Province, China
e-mail: ljh3832@163.com

Y. Wang
Machinery and Electronics Engineering, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun City 130033, China

establish local geometric feature matching from incomplete 3D data, this is a challenging work for modern 3D matching technology. A fine geometric descriptor can be obtained through manual operation, but this approach is based on a static geometric histogram which is unstable during the real-time scan [1–3], it also leads to inconsistent matching results.

In view of the above problems, we present a data driven neural network model in this paper. We get a robust local geometric feature descriptor by learning of incomplete 3D scanning data. A 3D convolution neural network model is established to match the local geometric features. The model is build based on the real scanning data of automobile casting, it combines the geometric feature representation with the metric function. In order to make the model being compatible with the 3D casting geometry, the 3D geometric features are encoded using the truncated distance function (TDF). The structural feature of TDF allows for the three-dimensional convolution and other kernel operations used in the model. TDF also supports the model to learn the geometric representation directly in 3D space. This geometric representation reduces the cumulative error during the matching process, it reduces the sensor noise while robustly aggregating multiple depth data [4]. In this paper, we use the existing RGB-D reconstruction algorithm to generate the correspondence label, the matching descriptor is trained on local geometry around the 3D Harris key. In the process of sensor scanning, the world space position of the feature points can be obtained from different camera perspectives, which allows our training model to automatically generate a large number of correspondence in real-time without manual operation. Because of the different occlusion angles of each camera, our model supports local to local ground matching, it allows for the robust corresponding between key points.

In this paper, the 3D reconstruction quality is neglected during the initial training process. The training sequence contains only a small number of matching features. However, once the reconstruction is done successfully, more and more key points can be generated even if the original feature matching fails before. Our method is based on data-driven 3D key descriptor, which is used to match local geometric features. Our descriptor performs well for 3D matching, it matches RGB-D scan data reliably and reconstructs 3D geometric structure of automobile castings accurately. For the incomplete sparse surface geometry, it is still possible to obtain globally consistent 3D reconstruction from the real automobile casting.

2 Related Work

Over the past decade, researchers have proposed a variety of manual geometric descriptors, including spin images, geometric histogram descriptor, etc. These methods have been integrated into the Point Cloud Library (PCL) framework [5]. The mainstream method of PCL is point feature histograms (PFH), it uses the surface normal and the curvature estimation to obtain the descriptor [6, 7]. Aiger et al. [8] proposed the use of four four-point congruent sets (4PCS) to obtain fragment alignment. Then, Mellado et al. extended the fragment alignment to super 4PCS [9]. While these methods perform well, it is still difficult to handle the interference of noise, low resolution, and incomplete scan data obtained from the depth sensor. Since the manual geometric descriptors are based on the descriptors derived from the static geometry histogram, the matching results are unstable and inconsistent during partial scanning. In order to improve the low precision problem of these geometric matching method, Choi et al. [10] delivered a robust method for the reconstruction of indoor scene, a special procedure is done for the optimization of unmatched region.

Due to the widespread use of large-scale data, it is possible to design a two-dimensional image descriptor based on a data-driven approach. For example, Zollhöfer learns a non-linear mapping from intensity segments to image feature descriptor in a predetermined metric range, typically using euclidean distance (ED) or mahalanobis distance (MD) Distance mapping [11]. Jain et al. [12] demonstrated that feature learning can be achieved by extending the learning feature descriptor to the feature comparison metrics. Recently, a deep convolution neutral network method has been used for learning of descriptors and related metrics for local 2D RGB blocks [13]. Based on these two-dimensional data-driven methods, this paper designs a three-dimensional geometric network structure, it owns an unified measurement for feature learning. It builds local geometric matching through the operation of three-dimensional data, later the 3D key points are corresponded. Based on the end-to-end

matching of feature metric, the geometric correspondence is robustly established between the incomplete, noise interfered 3D scan data.

In recent years, the convolution neural networks has been used for 3D depth data. For example, Wu et al. [14] proposed a 3D volume representation by depth date, he uses depth learning for 3D shape modeling, it shows that the three-dimensional feature can be learned from a large number of three-dimensional CAD models. In addition, most of the recent methods extract the depth learning features from the 3D data, these method are used for CAD modeling as well as object detection and classification [15, 16]. These methods extract the global features from the complete CAD model, on the other hand, we extract the local geometric features of automobile casting from its RGB-D scanning data, it allows for robust 3D reconstruction of incomplete automobile-casting geometries with noises.

3 System Overvies

A front-to-back reconstruction system framework is constructed in this paper, as shown in Fig. 1. The front-end involves the camera attitude estimation and the local fusion surface reconstruction, the back-end involves the attitude map optimization and the intensive depth map optimization, which is combined with the depth of the neural network model. It is trained to rebuild the label to ensure the consistency of global reconstruction. Based on GPU-accelerated fusion surface reconstruction and processing RGB-D data at 30 Hz camera frame rate, 45 M of learning data is processed per second. In this paper, the casting model is reconstructed into a dense 3D grid-based training framework, which does not depend on the original point cloud, which makes reconstruction more efficient. In this paper, the reconstructed labels are robust and can capture fine-grained features in real time on the surface of castings, and realize locally and globally consistent surface-intensive reconstruction.

Firstly, the RGB-D stream input is obtained from the depth sensor and a set of sparse corresponding term features are used to obtain coarse global alignment, and the alignment is corrected by optimizing the density and geometric consistency to calculate the paired scale of all input frames Feature transform. All SIFT keys are detached to match the previous frame, and the outliers is filtered out.

Secondly, in order to realize the global attitude alignment, the system uses the filtering frame to perform local to global pose optimization hierarchically. At the first level, consecutive n frames are localized to the local level. At the second level, all blocks are correlated and globally optimized. This algorithm generates blocks based on the currently visible frustum region (TSDF). And the posture alignment of the two stages is optimized according to the sparse corresponding term and the dense photometric geometric constraint. This hierarchical optimization strategy reduces the non-associated characteristic of each optimization step, which makes the algorithm suitable for casting structural reconstruction. The system uses the GPU nonlinear iterative solver to deal with the highly nonlinear optimization problem at two levels.

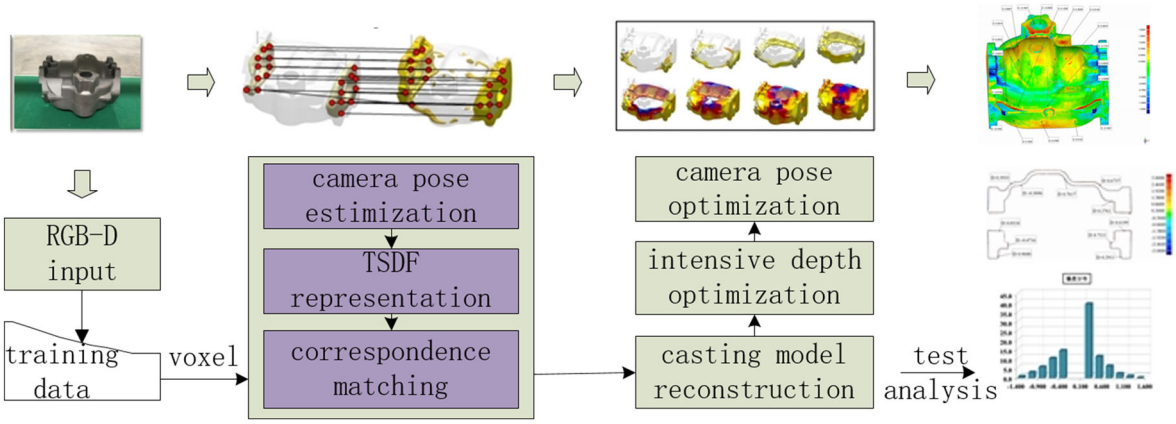

Fig. 1 Our reconstruction pipeline

Finally, the global representation is continuously updated according to the continuously changing global attitude (optimization), the RGB-D frames are quickly re-fused, and the old pose RGB-D images are removed using the anti-fade step, and the new gestures are re-merged into RGB-D image. With more RGB-D frames and sophisticated attitude estimation, the volume model is continually improved to ensure the quality of the reconstruction.

4 3D Geometric Representation

The core of 3D geometric matching is to establish a robust correspondence between geometric 'fragments'. The input data source for our neural network is the depth frame data captured from the depth sensor. These depth data are usually not aligned with any global coordinate system. Instead of using RGB color information, only the geometric information obtained from the depth sensor is used to complete the matching process. The geometric matching is achieved by comparing the sampling distance field around the key points. In this section, we'll show how to convert the scanned data into geometric fragments, how these fragments are represented as distance fields, and how to extract key points from these distance fields.

At first, consecutive deep frames N are fused to a distance field. When $N = 1$, each fragment contains only a single frame. As N increases, each fragment can integrate information from multiple depth frames. The more geometric information can smooth sensor noise and expand the fragment view. N should be maintained small enough so that local correspondence can be used to generate high quality fragments without accumulating excessive drift errors. In this paper, we use the iterative closest points (ICP) to align frames in fragment, where $N = (25,40)$. If a valid alignment can not be found, it means the lack of geometric feature, the fragment will be discarded. Our method is based on the truncated symbolic distance field (TSDF), the fragment fusion is obtained from the local alignment, the depth data is fused to the first frame voxel of the anchor fragment through the use of volume fusion algorithm [17, 18].

For each spatial voxel on automobile casting, we use $\mathbf{D}(v)$ to express the signed distance of voxel, $\mathbf{W}(v)$ represents the voxel weight, $d_i(v)$ represents the projection distance (along z-axis) between each voxel and the depth frame D_i. The volume fusion is updated for each voxel:

$$\begin{aligned} \mathbf{D}'(v) &= \frac{\mathbf{D}(v)\mathbf{W}(v) + w_i(v)d_i(v)}{\mathbf{W}(v) + w_i(v)}, \\ \mathbf{W}'(v) &= \mathbf{W}(v) + w_i(v) \end{aligned} \tag{1}$$

In this paper, the three-dimensional data is build upon the uniform sampling distance field which realize the kernel operation of the depth data, that is 3D convolution operation. Compared with coding for the 2D depth block, the voxelized 3D representation retains the spatial scale information of the real casting object. The voxel owns invariant characteristic when the projection and rotation are being done. By detecting and calculating the characteristics of the TSDF, we train the gradient-sensitive convolution neural network, while we limit the highest gradient on surface area to minimize the underlying kernel confusion. TSDF symbol is ignored here, no matter whether the free space is visible, and the highest gradient of distance field is concentrated on large surface area, rather than in the camera boundary of the frustum.

Based on the generated fragments and their TDFs, the key points in the fragments are then detected, their local TDF regions are extracted. Here we need to focus the descriptor in the geometric area of casting object. The granularity of the critical area (radius ≈ 1.5 cm) is suitable for typical automobile-casting reconstruction. Its local adaptive feature guarantees a certain range of context coverage, it also owns precise discrimination characteristics to capture enough geometric details. According to the 3D object retrieval method, the 3D Harris corner is used here to determine the key location. For each voxel adjacent to the mesh surface, the covariance matrix C with neighborhood normal n_i is determined according to the TDF gradient function, where the angular response: $r_i = \det(C_i) - 0.04 \cdot trace(C_i)$. For all of angular responses, non-maximum suppression is performed to reduce the number of samples,the iterations are applied to move the remaining samples to their local stable positions. The local area which is not observed by the frame of the key points is filtered out from neutral network. Then, we use the remaining key sparse collection and its local TDF volume as the descriptor input. We use a single thread on the INTEL Core i5/3.0 GHz CPU to implement a complete Harris key detection and

extraction process, the TDF volume geometry fragment occupies 512 × 512 × 1024 storage space, it only takes about a few seconds to complete the entire detection process. Figure 1 shows the key points detected on the fragment. The green point in Fig. 1 identifies the unobserved key points, these local areas will be discarded, because they are not covered by enough frames. The well-covered red keys and their local regions are remained for training the 3D matching descriptor. In Fig. 2, the fragment pairs of the cast object are given from six different angles, their feature correspondence keeps in match with neural network. In this paper, we use these matching sparse geometric features to obtain 3D casting reconstruction. In this paper, we use these matching sparse geometric features to obtain 3D reconstruction.

5 Training Labels

The depth learning algorithm requires a large number of training data with real ground tags. Although it is easier to obtain data through a depth sensor, it is often necessary to obtain a large number of manual operations to obtain interrelated labels. In order to learn the correspondence between the key points, the local tag will involve a key point correspondence between millions of geometric fragments, which can not be achieved by manual operation. The existing RGB-D reconstruction algorithm can accurately align and fuse the depth frame, we uses Lin's reconstruction results to generate large-scale point-to-point corresponding labels [19]. Due to the use of RGB-D scanning, the same casting is recorded several times to ensure a different camera trajectory from different viewpoints.

First, in order to achieve a globally consistent reconstruction, we use state-of-the-art sparse–dense constraint adjustment taking into account both RGB and depth data. The TDF voxel pair between the different views is then sampled to generate the tagged training data, which in turn traces the descriptor and measures the network. The key of our approach is to weaken the dependence of the geometric features during the original reconstruction process. Only a few sparse features can be used to reconstruct the training sequence, and even if the original feature matching fails, once the reconstruction is successful, more key point correspondence can also be generated.

In this paper, the training labels are generated from the following real car cast objects. (1) Real Data Set. It contains RGB-D frames that are tracked on six complex car cast surfaces. It performs "local live" tracking using ICP and frame-model alignment. The training fragment uses the five sequences in these scenarios, the test clip uses the 6th sequence. (2) Synthetic data set: It contains 12 different casting data from camera by tracking RGB-D frame data. The training fragments in this paper use 10 sequences in these objects, and the test clip contains two other castings [20, 21].

These datasets are chosen because they provide different perspectives of the same cast from a variety of different perspectives. It provides a distributed real sample of possible scans for any given key. Since the

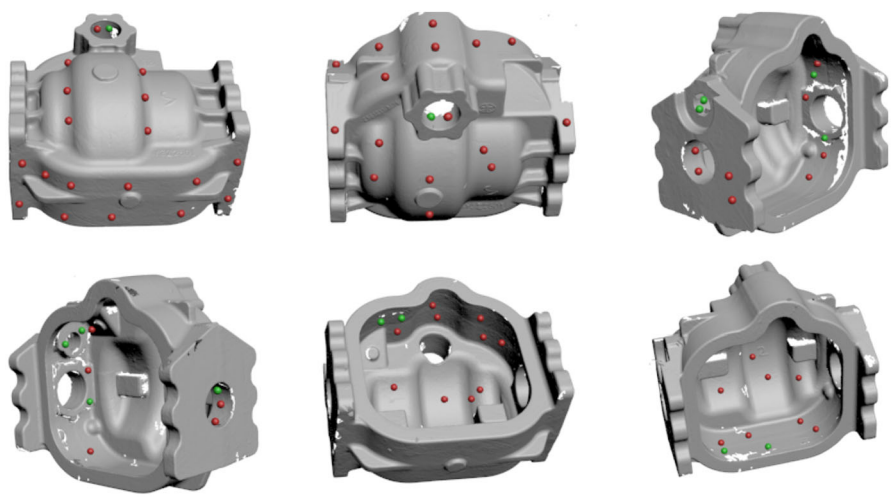

Fig. 2 Fragment detected by the matching key. The *green points* represent the unobserved key points, the uncovered local areas with *green keys* will be discarded. The well-covered *red keys* and their local regions are remained for training the 3D matching descriptor. (Color figure online)

local key points are oriented according to the camera frame, the training data from these data sets contains the correspondence between the key points, and the local volume covers attributes that reflect the change and occlusion of the viewpoint. These complex features are combined by training descriptor to simulate local 3D data for automobile castings.

Wherein each fused fragment contains 30 frames, the fragment dataset contains 1600 training fragments and 500 test fragments. Typically each fragment contains 200 to 500 Harris keys. We use the pseudo-ground conditions provided from the training dataset to correlate their world coordinates, while the ground truth correspondence is established between the key points of the same cast. A key point within 1.5 cm is marked as 'match' and a key point of more than 1.5 cm is marked as 'mismatched'. The 1.5 cm threshold for search radius is set to handle small alignment errors coming from outside ground conditions without being affected by the small translation differences between Harris key points.

6 3D Matching Neural Network

The 3D matching network model is a neural network architecture with uniform depth. The matching descriptor is composed of two core components: the first is feature network, the local 3D TDF volume is mapped to the high-density 3D feature representation network. The second is a metric network that maps feature pairs to similar values through a set of fully connected inner layers. Figure 3 shows the structure of network in this paper.

First, for each query key, the local TDF volume is cut out from the geometry fragment. These volumes are mapped entirely to feature descriptors containing 2048 elements through the feature network. Then, these feature vectors are connected to the metric network. The metric network divides the two points into 'match' and 'mismatch'.

The feature network of the 3D matching model maps the 3D local area with key points to the feature descriptor function. In this paper, the local radius of the key point is set to 1.5 cm, the input structure of the feature network is 31 × 31 × 31 voxel TDF (voxel size is 0.1 cm, truncation distance is 1.5 cm). The feature representation is a 2048 dimension eigenvector, feature network includes four non-linear

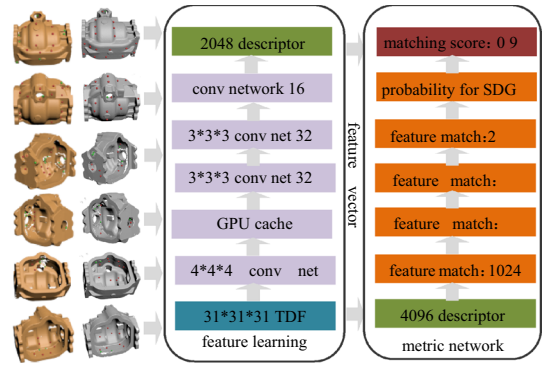

Fig. 3 3D matching network. Feature network includes four non-linear convolution layers. The metric network is a nonlinear matching function, which compares the two feature representation to determine the corresponding relationship between the two key points

convolution layers. Since the size of the initial input volume is small, our algorithm performs a single layer merge. The size of the kernel and the number of filters are shown in Fig. 2. The last concatenation layer of the feature network is used to determine the dimensions of the feature representation and to prevent excessive breakdown at the network level. In this paper, the core of the metric network is a nonlinear matching function, which compares the two feature representation to determine the corresponding relationship between the two key points. The input of the network is two eigenvector, while the output is a single confidence value between 0 and 1, which is used to simulate the similarity between the keys, where 1 is "match" and 0 is "mismatch".

The metric network consists of several completely connected layers, the last layer is estimated by the probability of grid. Its two values represent the probability matching of two features whether matching or mismatching. In this paper, the stochastic gradient descent (SGD) training set is used to define the cross entropy error as:

$$E = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \log(\hat{y}_i)] \quad (2)$$

in which, y_i is a binary tag ('match' or 'mismatch') for input data x_i, $\hat{y}_i$ is the value of the network output from the estimated layer. In order to estimate the matching performance of the metric network, the metric network is replaced by a single contrast loss layer, and the bottleneck feature is compared using the euclidean

distance (ED). This replacement reduces key point matching performance, but it is superior to manual descriptor methods.

In order to detect the types of information captured by neural networks, we randomly extract 2000 key volumes from the test casting, we find out their two-dimensional embedding of 2048 eigenvectors. For each key TDF volume, its grid is generated and positioned in the embedded location. For each key grid, it is colored using a normalized three-dimensional embedding of the eigenvector. The overall layout of the embedding indicates that the feature network is able to aggregate different types of local casting geometries, such as edges, planes, and corners.

Due to the great changes in training data, a large number of training iterations is needed to ensure convergence. Our method supports the N-dimensional convolution neural network depth learning framework. In order to train the network, all the layers (standard deviation is 0.01, initial deviation is 0) are randomly initialized according to the zero Gaussian plotting weights. In order to adjust the training bias, the number of matched volumes corresponds with the number of unmatched volumes. The basic learning rate is set to 0.01, the learning rate is increased by 0.99 factors per iteration, each iteration contains 2000 sub-iterations. SGD runs 1.3 million iterations with a momentum of 0.9, parameter attenuation is 0.0005.

7 Experimental Results and Analysis

In this paper, both the real automobile casting scene and synthetic data are used to test the neural network model. The experimental parameters are set as follows: operating system is Windows 8.1; Development language is C++ by Microsoft Visual Studio 2014; We use the Direct3D 11 as 3D graphics programming interface; CPU type/frequency is INTEL Core i5/3.0 GHz with 8G RAM; The type of graphical processor is NVIDIA GTX 960; We capture depth data through Asus Xtion Pro sensor, its RGB-D streaming rate is 30 Hz wiht 640 × 480 depth resolution. In Sect. 7.1, several test results are presented to evaluate the performance of our method. In Sect. 7.2, our method is compared with several geometric matching methods. A complex casting reconstruction test is given in Sect. 7.3, and the different training scenarios are reconstructed for test.

7.1 Quantitative Analysis

For different scans of the same casting, the performance of 3D geometric descriptor is evaluated by testing whether the two points match in correspondence. The ratio between matching and non-matching is 1:1 for a local volume that contains 25,000 pairs of points, a small number of key volumes is detected by corresponding labels as “matched” or “unmatched”. Key correspondence is generated from different scans of the same casting. We construct two test datasets: Harris and random data sets. In the Harris data set, the points are randomly selected from the set of Harris keys containing at least one correspondence; in the random data set, the points are randomly selected from the fragments. Several matching algorithms are tested on both data sets. For spin images and fast point feature histograms, we use the adjustment parameters provided in the point cloud library as a benchmark. These methods operate directly on the point cloud of the fragment, it does not include any additional information from the signed distance field.

Figure 4 shows the accuracy and performance of several geometric descriptors on the matching confidence threshold (or distance threshold). Our descriptor performs well by retaining 81.6% accuracy at 95% closed loop. The matching threshold for each algorithm is calculated by matching the optimal ratio of the volume, while the ratio is changed in the evaluation data set by random sampling, and the accuracy of the descriptor keeps relative to that ratio. The experiment shows that manually produced descriptors are poor in performance in the face of noise and data loss. Our method in this condition can maintain a more consistent accuracy. Table 1 gives the quantitative results of the precision reconstruction. Our method owns a higher precision compared with other manual geometric matching methods.

Under the different scan in the same casting, we test the matching ability of our neural network for 3D key points. Fragments of two RGBD frame sequences from the same casting are fused for reconstruction. These two sequences are called query sequences and sample sequences respectively. The two sequences are related to each other in world space. In each of the two

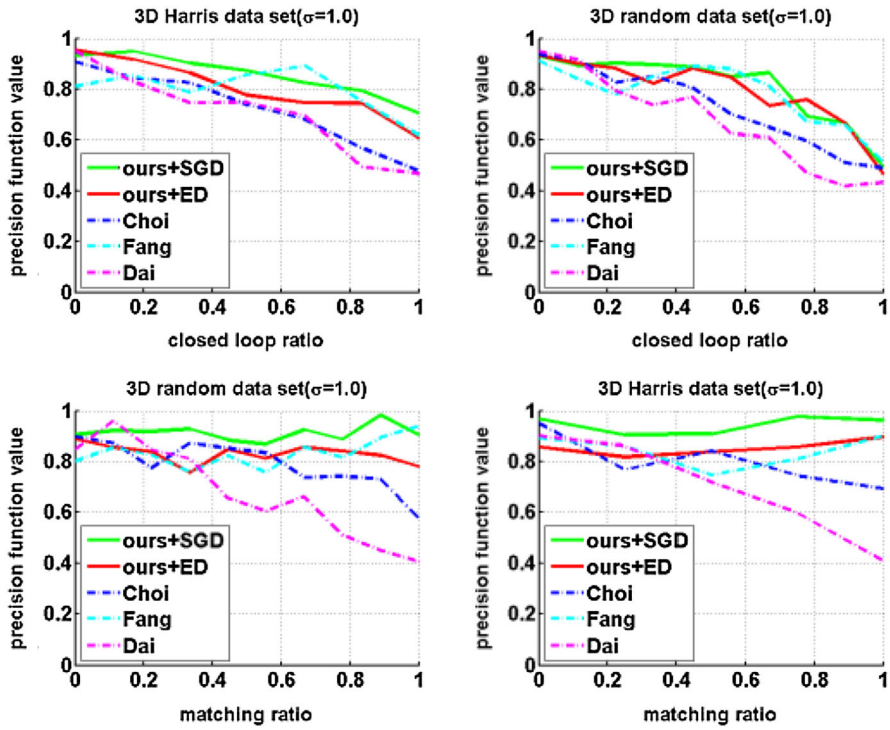

Fig. 4 Comparison of several algorithms for the matching accuracy. The graph shows the performance of several state-of-art geometric descriptors on the matching confidence threshold

Table 1 Reconstruction accuracy percentage

Method	Harris data set (%)	Random data set (%)
Dai	60	55
Fang	78	70
Choi	75	68
Ours	81.6	75

Our method is compared with other methods for reconstruction precision ratio

sequences, a set of Harris key points are calculated, their local TDF volume is extracted. In the world space, all the Harris key points within 1.5 cm are considered to be true to the ground. For the scanned camera frame, each key point and its local TDF volume are anchored, the key volume here is oriented according to a variety of different camera angles.

For each key in the query sequence, our model checks whether the descriptor is able to find the true ground value from the sample sequence including n-level similarity score. When $n = 1$, the percentage of key points of the highest similarity are returned by comparison of the key points in query sequence. Figure 5 shows the percentage of key points that have been successfully found in the first n matches, where n

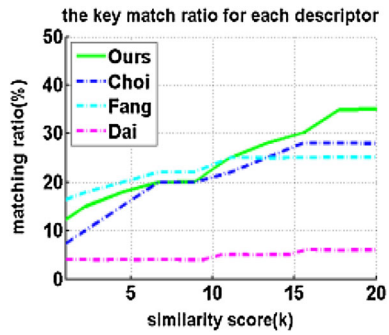

Fig. 5 Descriptor matching ratio for several algorithms. The percentage of detected key points from the first n matches are shown in graph, where n varies from 1 to 20

varies from 1 to 20. When $n = 5$, 20% of the key points finds the right match, while Dai method can only find 4%. Figure 6 shows the three-level key points that match the two fragments under different scans. The gray fusion graph represents the initial sample, using 7 test fragments to detect the Harris key, and the yellow segment represents the first three matching Harris key points. Our method matches the detection fragments from different camera angles, the incremental fusion is performed in better effect for initial samples, the matching result is well.

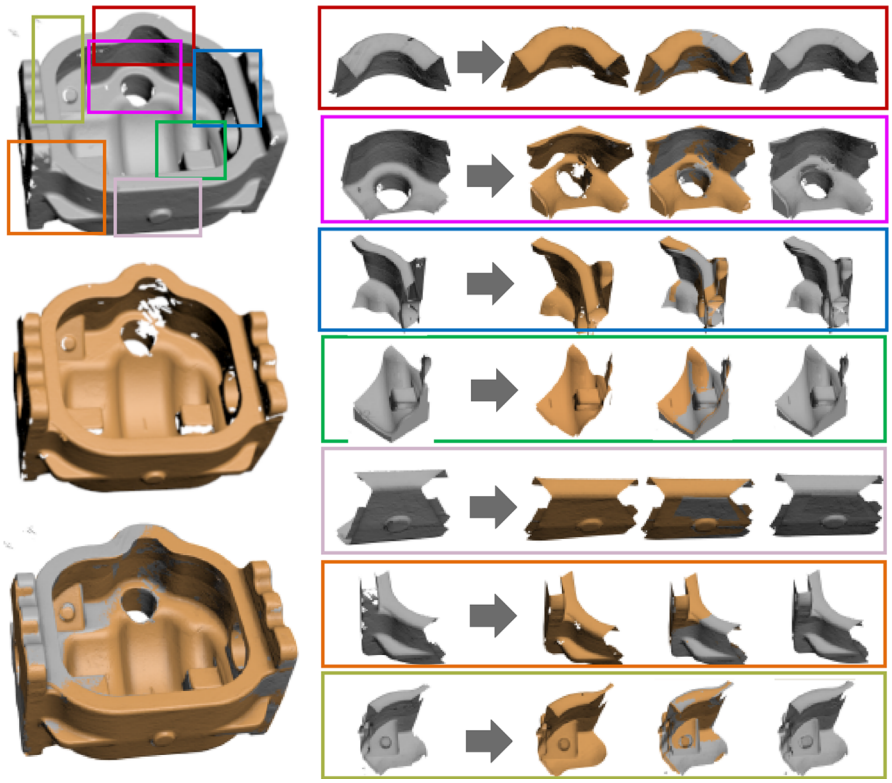

Fig. 6 The matching process of our descriptors. We select seven overlapping fragments to detect the Harris key. The *gray* fusion graph represents the initial sample, the *purple* fragment represents the first three matching Harris key points. Our method matches the detection fragments from different camera angles. (Color figure online)

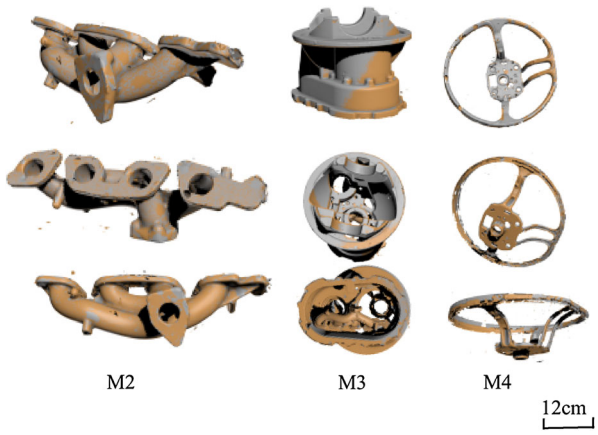

Fig. 7 The matching results of our method. The *gray part* shows the initial fusion result, the *yellow* parts shows the fusion fragments of our descriptor. It performs well for 3D geometric matching. (Color figure online)

7.2 Qualitative Analysis

The performance of our descriptor is tested in this section, fragment fusion is performed after PCL, and features are computed on a set of Harris key points, as shown in Fig. 7. First, each Harris key is mapped to the n correspondence with the matching network. Then for each iteration, we randomly select the three correspondence to estimate the rigid transformation. The final transformation is internal correspondence with highest number n, correspondence keeps alignment within 1.5 cm, matching score is of at least 0.5.

In Fig. 8, it shows the test results of our proposed method compared with the geometric matching method proposed by Choi et al. The alignment of our method performs well, when τ is small, our method is more suitable for the matching of local

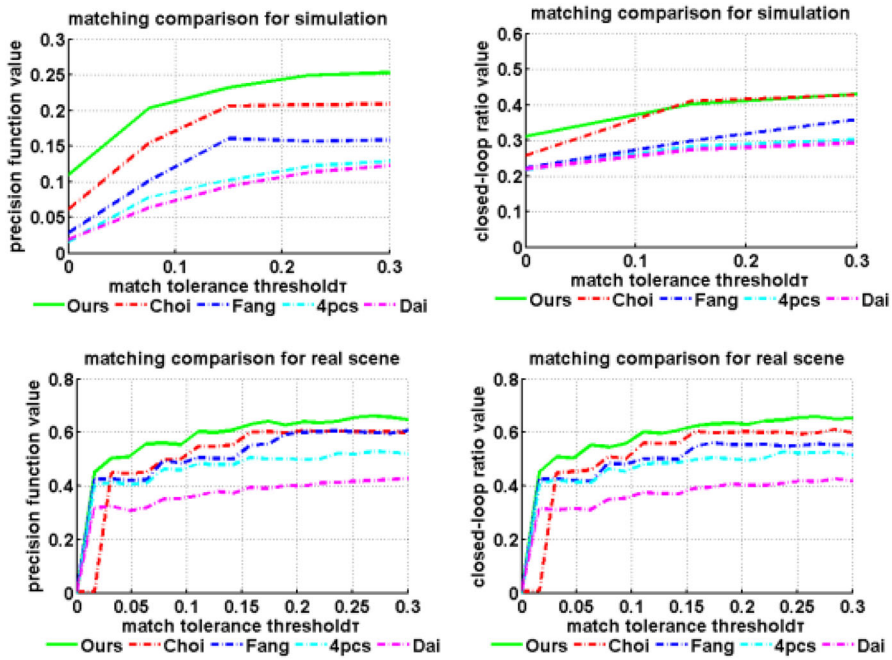

Fig. 8 Comparison of the matching accuracy between several different algorithms. Our method performs well for feature matching, when τ is small, our method is more suitable for the matching of local details

details. When $\tau = 0.15$, Choi's 3D matching accuracy is better, the maximum accuracy is 60.0%, the average closed-loop rate is 60.6%. For the accuracy and closed-loop results of the matching tolerance threshold τ in the real object, the comparison time for each fragment is limited to 10 s. In this paper, when $\tau = 0.2$, the matching result is better for complex structured casting with 92.2% accuracy and 74.0% closed loop rate.

7.3 Reconstruction Error Analysis

The difficulty of 3D casting reconstruction lies in the closed loop, that is, we need to form a corresponding relationship between the same position from different perspectives. Both color and depth information provide data channels that can be used to detect these correspondences. However, color-based descriptor (such as SIFT) often fails to find the correct correspondence when there is a wide baseline view change or a violent light difference. Figure 9 shows that in the complex closed-loop scenario, our matching network uses geometric information to align these loop closure fragments more accurately. Our method is compatible with sparse RGB features. When there are insufficient geometric information in the fragment, the combination of SIFT with our method can improve the alignment accuracy.

The automobile casting model M1 and M4 are tested and analyzed to calculate the 3D deviation. The results show that the average reconstruction error of our method is 0.3943/0.4933, the maximum and minimum critical value is ±1.5 mm. Standard deviation is about 0.5772, the deviation is of less than 0.3000 for 40.2 percent 3D point. Our neural network can be combined with sparse RGB features. When there are incomplete geometric information in the fragment, the combination of SIFT with our method can improve the alignment accuracy.

8 Conclusion

Based on data-driven strategy, we construct a geometric matching network, a 3D geometric matching neural network model is proposed. In this paper, we study the real data of automobile castings, training the geometric descriptor, we testify its robustness in key point matching, fragment alignment and scene reconstruction. It is believed that the research on RGB-D feature matching will continue be prosperous in many aspects, for example, the mapping of correspondence between different types of sensors ignoring time and lighting conditions, and a more robust data-driven strategy for 3D reconstruction.

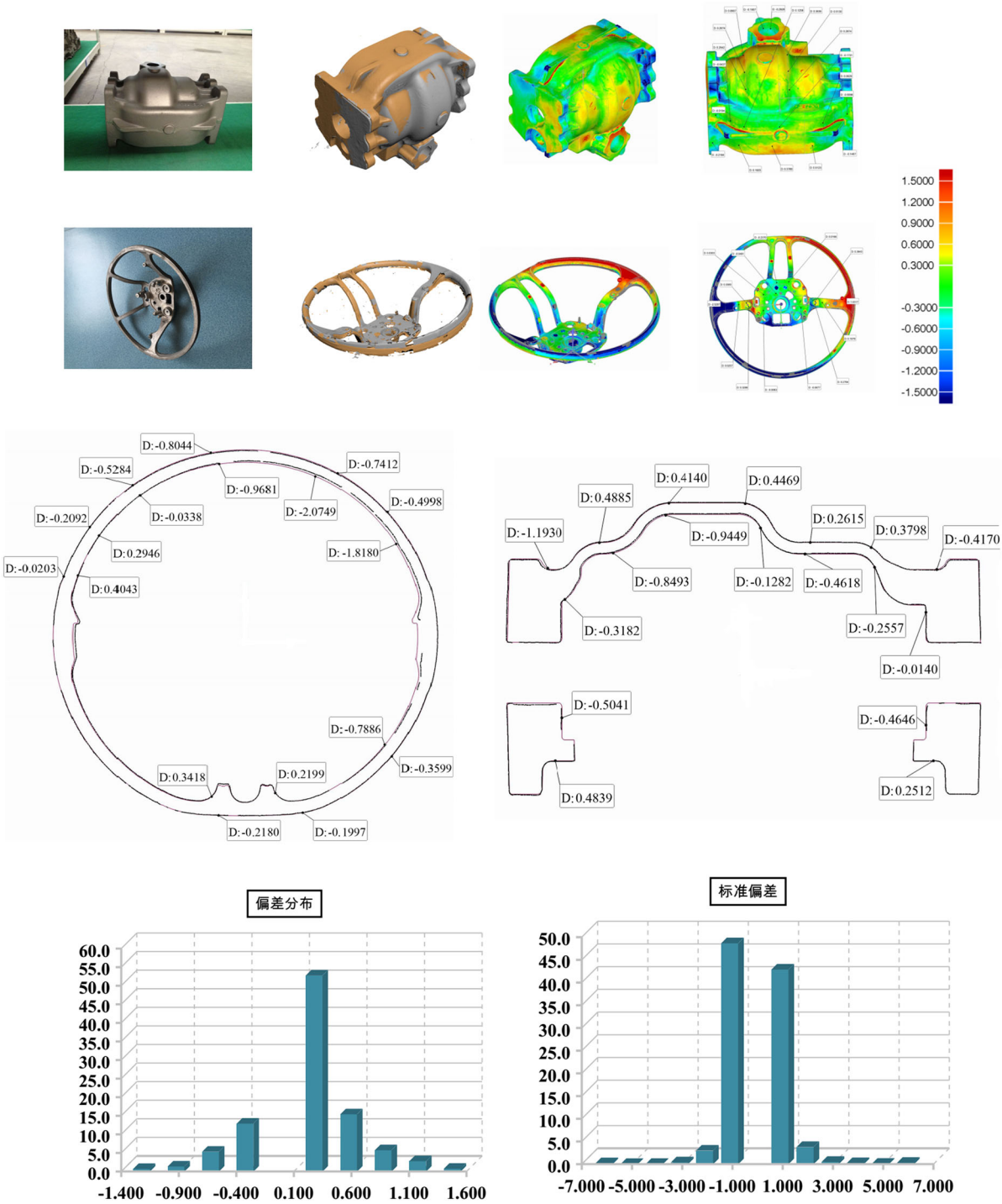

Fig. 9 Comparison of reconstruction results. The *second line* graphs show the fusion process of our matching network. When there are no sufficient geometric information in the fragment, the combination of SIFT with our method can improve the alignment accuracy

Acknowledgements This work was supported by National High-tech R&D Program (Grant No. 2014AA7031010B); Science and Technology Project of The thirteenth Five-Year Plan (Grant No. 2016345).

References

1. Weise, T., Wismer, T., Leibe, B., et al. (2009). In-hand scanning with online loop closure. In *IEEE international conference on computer vision workshops* (pp. 1630–1637).
2. Henry, P., Krainin, M., Herbst, E., et al. (2012). RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. *International Journal of Robotics Research, 31*(5), 647–663.
3. Keller, M., Lefloch, D., Lambers, M., et al. (2013). Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *International conference on 3d vision-3dv* (Vol. 8768(2), pp. 1–8).
4. Curless, B., & Levoy, M. (1996). A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on computer graphics and interactive techniques* (pp. 303–312). ACM.
5. Aldoma, A., Marton, Z. C., Tombari, F., et al. (2012). Tutorial: Point cloud library: Three-dimensional object recognition and 6 DOF pose estimation. *Robotics & Automation Magazine IEEE, 19*(3), 80–91.
6. Whelan, T., Leutenegger, S., Salas-Moreno, R. F., et al. (2015). ElasticFusion: Dense SLAM without a pose graph. Robotics: Science and Systems (RSS), Rome, Italy.
7. Meilland, M., & Comport, A. (2013). On unifying key-frame and voxel-based dense visual slam at large scales. In *IEEE/RSJ international conference on intelligent robots & systems* (Vol. 8215(2), pp. 3677–3683).
8. Aiger, D., Mitra, N. J., & Cohen, D. (2008). 4-points congruent sets for robust pairwise surface registration. *ACM Transactions on Graphics, ACM, 27*(3), 85.
9. Mellado, N., Aiger, D., & Mitra, N. J. (2014). Super 4pcs fast global point cloud registration via smart indexing. *Computer Graphics Forum, 33*(5), 25–215.
10. Choi, S., Zhou, Q. Y., Koltun, V. (2015). Robust reconstruction of indoor scenes. In *IEEE conference on computer vision and pattern recognition* (pp. 5556–5565).
11. Zollhöfer, M., Thies, J., Colaianni, M., et al. (2014). Interactive model-based reconstruction of the human head using an RGB-D sensor. *Computer Animation & Virtual Worlds, 25*(25), 213–222.
12. Jain, P., Kuli, B., Davis, J. V., et al. (2009). Metric and kernel learning using a linear transformation. *Journal of Machine Learning Research, 13*(1), 519–547.
13. Han, X., Leung, T., Jia, Y., et al. (2015). Match net: Unifying feature and metric learning for patch based matching. In *IEEE conference on computer vision and pattern recognition* (pp. 3279–3286).
14. Wu, Z., Song, S., Khosla, A., et al. (2015). 3D ShapeNets: A deep representation for volumetric shapes. In *Computer vision and pattern recognition* (pp. 1912–1920).
15. Fang, Y., Xie, J., Dai, G., et al. (2015). 3d deep shape descriptor. In *IEEE conference on computer vision and pattern recognition* (pp. 2319–2328).
16. Dai, A., Nießner, M., Zollhöfer, M., et al. (2016). BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface re-integration. arXiv preprint.
17. Qu, Y., Liu, Z., Jiang, Y., et al. (2017). Self-adaptative variable-metric feature point extraction method. *Editorial Office of Optics and Precision Engineering, 25*(1), 188–197. **(in Chinese)**.
18. Liu, Y., Wang, C., Gao, N., et al. (2017). Point cloud adaptive simplification of feature extraction. *Editorial Office of Optics and Precision Engineering, 25*(1), 245–254. **(in Chinese)**.
19. Lin, J. H., Wang, Y. J., & Sun, H. H. (2017). A, Feature-adaptive subdivision method for real-time 3D reconstruction of repeated topology surfaces. *3D Research, 8*(1), 1–16.
20. Sturm, J., Engelhard, N., Endres, F., et al. (2012). A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE international conference on intelligent robots and systems* (pp 573–580).
21. Zhang, L., Wang, Y., Sun, H., et al. (2016). Adaptive scale object tracking with kernelized correlation filters. *Editorial Office of Optics and Precision Engineering, 24*(2), 448–459. **(in Chinese)**.