A reliable quasi-dense corresponding points for structure from motion

Jangseok Oh[1], Hyunggil Hong[1], Yongjun Cho[1], Haeyong Yun[1], Kap-Ho Seo[1,2], Hochul Kim[3], Mingi Kim[4], and Onseok Lee[5*]

[1]Agriculture Robotics & Automation Research Center, Korea Institute of Robotics and Technology Convergence, Andong City, Republic of Korea
[e-mail: dueleldi@kiro.re.kr, honghg@kiro.re.kr, cyj@kiro.re.kr, hyyun@kiro.re.kr, neoworld@kiro.re.kr]
[2] Department of Mechanical Engineering, Pohang University of Science and Technology, Pohang City, Republic of Korea
[e-mail: neoworld74@gmail.com]
[3] Department of Radiological Science, Eulji University, Seongnam City, Republic of Korea
[e-mail: tiger1005@eulji.ac.kr]
[4]Department of Electronic and Information Engineering, Korea University, Seoul, Republic of Korea
[e-mail: mgkim@korea.ac.kr]
[5]Department of Medical IT Engineering, College of Medical Sciences, Soonchunhyang University, Asan City, Republic of Korea
[e-mail: leeos@sch.ac.kr]
*Corresponding author: Onseok Lee

Received April 9, 2020; revised July 13, 2020; accepted August 15, 2020; published September 30, 2020

Abstract

A three-dimensional (3D) reconstruction is an important research area in computer vision. The ability to detect and match features across multiple views of a scene is a critical initial step. The tracking matrix W obtained from a 3D reconstruction can be applied to structure from motion (SFM) algorithms for 3D modeling. We often fail to generate an acceptable number of features when processing face or medical images because such images typically contain large homogeneous regions with minimal variation in intensity. In this study, we seek to locate sufficient matching points not only in general images but also in face and medical images, where it is difficult to determine the feature points. The algorithm is implemented on an adaptive threshold value, a scale invariant feature transform (SIFT), affine SIFT, speeded up robust features (SURF), and affine SURF. By applying the algorithm to face and general images and studying the geometric errors, we can achieve quasi-dense matching points that satisfy well-functioning geometric constraints. We also demonstrate a 3D reconstruction with a respectable performance by applying a column space fitting algorithm, which is an SFM algorithm.

***Keywords*:** scale-invariant feature transform, speeded up robust features, structure from motion, column space fitting, affine-model

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through Advanced Production Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) (317072-04, 318072-03) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2018R1D1A1B07049072, 2019R1F1A1058827).

http://doi.org/10.3837/tiis.2020.09.012 **ISSN : 1976-7277**

1. Introduction

Correct matches can be used to perform many different computer vision tasks such as an estimation of a three-dimensional (3D) reconstruction, camera motion, recognition, and simultaneous localization and mapping (SLAM). To generate a mosaic of images, it is important to achieve an accurate match [1]. Because binocular stereo vision imitates human eyes and can fuse two images into a stereo mapping image, studies on a 3D reconstruction using this approach have also been conducted [2].

To build an accurate 3D model, it is necessary to detect and match a considerable number of reliable features across video frames. Scale invariant feature transform (SIFT) [3] and speeded up robust features (SURF) [4], the best-known point extraction algorithms, are state-of-the-art interest point detectors. In addition, a novel method has been developed based on SURF points for the detection of moving targets [5]. Deep learning has also been used in recent studies on the feature matching problem. Zhang *et al*. [6] proposed the use of a graph neural network to exploit the global structure of a graph for transforming weak local geometric features at different points into rich local features for dealing with the geometric feature matching problem. Ufer *et al*. [7] introduced an efficient Markov random field (MRF) framework utilizing deep feature pyramids and convolutional activation guided feature selection for semantic matching. In a previous study, we focused on face recognition by applying SIFT using an adaptive threshold [8]. We obtained acceptable matching points by combining a robust principle component analysis [9] and adaptive sparse code shrinkage [10] in endoscopic images with noise.

Our goal is to conduct an accurate 3D reconstruction. One of the most frequently used structure from motion (SFM) algorithms is column space fitting (CSF) [11], which produces a sparse reconstruction from the matching and tracking methods applied. Another algorithm used for a sparse reconstruction is incremental SFM [11–14]. A dense matching process using patch-based multi-view stereo [15] is included to achieve a dense reconstruction. To improve the accuracy, a complex sparse bundle adjustment process [16] is also required. The proposed algorithm achieves a quasi-dense reconstruction without additional steps, using quasi-dense matching, and has a high level of accuracy. To accomplish this, we apply an affine model [17, 18]. A different tilt value from affine SIFT (ASIFT) [17] is used in fully affine invariant SURF [18], and this value is standardized with that developed by Morel *et al*. [17], the result of which is called affine SURF (ASURF).

2. Method

2.1 Affine models

Morel *et al*. [17] proposed a robust ASIFT algorithm that enhances the existing SIFT matching method to consider a change in the angle of the camera axis. As indicated in (1) and (2), a geometric interpretation of the six affine parameters is considered. The existing SIFT extracts the feature points by considering only four of the six parameters (scale, λ; rotation, ψ; x-axis translation, e; and y-axis translation, f) to draw the regional characteristics between images. Consequently, scale invariance occurs. However, as indicated in the equation, fully affine invariance occurs only when all six parameters are considered. We conduct a simulation for each case where affine distortion from a camera axis change is possible. Here, A indicates

a planar projective transform (homography). In an affine map, A is the determinant, λ represents the change in size, ψ is the rotational change, and the image distortion is dependent on the latitude (ψ) and longitude (ϕ).

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} \tag{1}$$

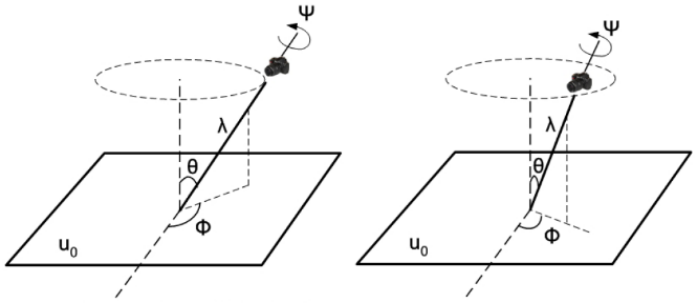
$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \tag{2}$$

$$t = 1/|\cos\theta|,\ t > 1 \leftrightarrow \theta \in [0°, 90°] \tag{3}$$

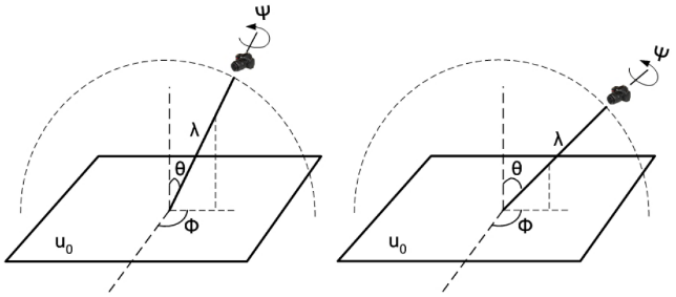
where λ is the zoom parameter, ψ is the rotation angle of the camera around the optical axis, θ is the latitude angle between the optical axis and the normal to the image plane, ϕ is the longitude angle between the optical axis and a fixed vertical plane, and $\begin{pmatrix} e \\ f \end{pmatrix}$ are the translation. As indicated in (2) and (3), for the simulation, the latitude angle uses tilt t. We can draw the feature points by producing images that consider the change in the longitude and latitude angles. **Fig. 1** illustrates the two main parameters (θ, ϕ) in the affine image deformation. Here, u_0 is a frontal view of a flat object. According to Morel *et al.* [17], we must consider all possible images resulting from the changes in latitude and longitude. However, in our study, we only sampled images that consider the changes listed in **Table 1**. The sampling location can be increased or decreased if required, as described in the studies by Pang *et al.* [18]. The feature points can then be determined by applying the SIFT algorithm to each image obtained. The location of the feature point found in the sampled image is calculated into the location of the original image. For our study, the model names are as listed in **Table 1**. For the original, θ is zero, and hence there is no sampled image. The number of simulated images is determined as follows based on the longitude angle:

$$\Delta\phi = \frac{72°}{t} \tag{4}$$

Thus, ASIFT is called Aff1-SIFT to Aff5-SIFT, depending on the detailed location of the sampling. As shown in **Table 1**, 4 sampled images are applied for the longitude angle of 45°, which increases to 15 images for the maximum longitude angle of 80°. For Aff5-SIFT, the numbers of original and sampled images of the different longitude and longitude angles are combined to measure a sufficient number of feature points from 45 images. Aff1-SURF to Aff5-SURF use a similar method. Furukawa *et al.* [15] used both the difference-of-Gaussian (DoG) point detector and Harris corner detector for dense matching because they allow the determination of mutually complementary feature points. SIFT is a DoG-based blob detector, and SURF is a Hessian-based blob detector. Hessian-based detectors are more stable and repeatable than their Harris-based counterparts. Thus, we also use a combined SIFT and SURF (CSS). As illustrated in the next section, this also allows the drawing of mutually complementary feature points. We used the combined ASIFT and ASURF for dense matching, which is our main goal. That is, Aff5-CSS is the feature point extraction for our dense reconstruction.



(a) Changes in Longitude



(b) Changes in Latitude

Fig. 1. Geometric interpretation

Table 1. Models with fixed tilts

Model	Original	Aff1	Aff2	Aff3	Aff4	Aff5
t (tilt)	1	$\sqrt{2}$	2	$2\sqrt{2}$	4	$4\sqrt{2}$
θ	0º	45º	60º	70º	75º	80º
# simulated images	1	4	6	8	11	15

2.2 Adaptive thresholds

2.2.1 SIFT

The SIFT algorithm sequence is as follows:

1. Scale-space extrema detection,
2. Keypoint localization,
3. Orientation assignment,
4. Keypoint descriptor.

The SIFT algorithm can be divided into four main steps. To begin, the candidate feature points are extracted in the scale space. The safety of the candidate feature points is then examined, and the location of the stable feature points are set to detailed locations. Next, the reference direction for the corrected feature points is assigned. Finally, a feature point

technician within the local image of the peripheral region, based on the reference direction, is formed.

To increase the level of safety during the second process, the candidate feature points are eliminated using the contrast threshold value; the original SIFT algorithm uses a fixed value of 0.02. However, to apply this algorithm based on different application areas, we must find an appropriate value. To resolve this, instead of using a fixed value, we can adapt using the histogram of $|D(X)|$. By selecting the highest histogram of $|D(X)|$, we can extract sufficient and appropriate feature points for the image regardless of the image properties.

2.2.2 SURF

Similar to SIFT, the SURF method is an algorithm that extracts feature points using regional characteristics. The SURF descriptor forms a smaller detailed 4 × 4 square sub-region area focused around the feature point. The descriptor records 2, 4, and 8 characteristics in each detailed area and creates a descriptor vector in 32, 64, and 128 dimensions. The image matching speed is fast because internally it uses lower-dimensional descriptive data compared to SIFT and is based on the Hessian method, which allows it to quickly compute stable feature points. SURF, which has a faster matching speed than SIFT, uses a fixed Hessian response threshold value of 0.0002 and adjusts the number of extracted feature points. As mentioned above, choosing the highest value of the histogram of the Hessian response threshold makes it an adaptive algorithm.

2.3 Epipolar geometry

2.3.1 Fundamental matrix (F)

Useful geometric information for reconstructing the 3D information can be extracted from the fundamental matrix. The fundamental matrix F satisfies the following condition:

$$x'^{T}Fx = 0 \tag{5}$$

where $x \leftrightarrow x'$ is any pair of corresponding feature points in the two images. Corresponding points that satisfy (5) are correctly matched.

2.3.2 Geometric error

We must minimize the geometric error that will otherwise negatively influence the reconstruction. The geometric error (d, d') is illustrated in **Fig. 2**.

The minimization of the geometric error involves minimizing the following function:

$$C(x, x') = d(x,\hat{x})^2 + d(x',\hat{x}')^2 \text{ subject to } x'^{T}Fx=0 \tag{6}$$

where d(*,*) is the Euclidean distance between points, and $\hat{x} \leftrightarrow \hat{x}'$ is any pair of corresponding points satisfied by the epipolar geometry. This cost function can be minimized using a numerical minimization method such as the Levenberg–Marquardt method. In our experiments, we use a sixth-order approximation for the optimal triangulation. The epipolar geometry is described in detail by Hartley and Zisserman [19]. We use the value of the geometric error to calculate the accuracy of the matching point.

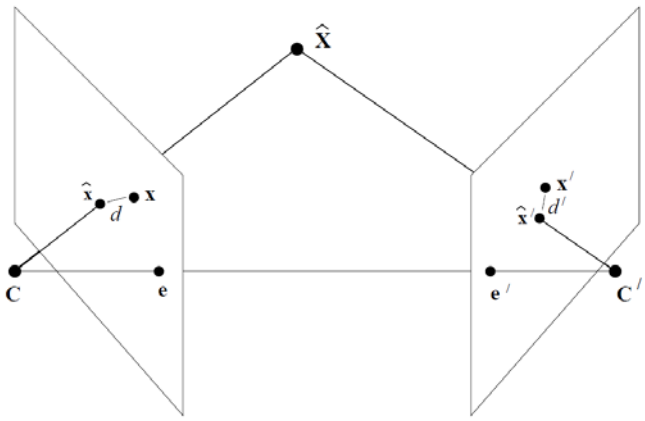


Fig. 2. Minimization of geometric error (d, d′)

2.4 Matching points

2.4.1 Initial matching

We follow Lowe's method [3]. To begin, we apply keypoint matching. The points with the closest distance become candidate keypoints. For a more accurate matching, we calculate the points with the closest and second-closest distances. If the distance ratio between these two points is greater than 0.8, the points are eliminated.

2.4.2 Remove outliers with epipolar geometry

We then conduct the following procedure to calculate a meaningful matching point.

(1) Solution for affine parameters

An affine transformation correctly accounts for a 3D rotation of a planar surface under an orthographic projection [3]. The matching points that do not fit this model are eliminated by calculating the six parameters, as shown in (7).

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{7}$$

For this calculation, at least three matching points are needed. To remove the outliers, we apply the calculations using the Random sample consensus (RANSAC) method. Further, these

parameters can be used for object recognition between the image and model.

(2) Computer fundamental matrix

We then determine the matching points that satisfy (5). This is also calculated using the RANSAC method to eliminate outliers. For this calculation, at least eight matching points are needed.

(3) Computer geometric error

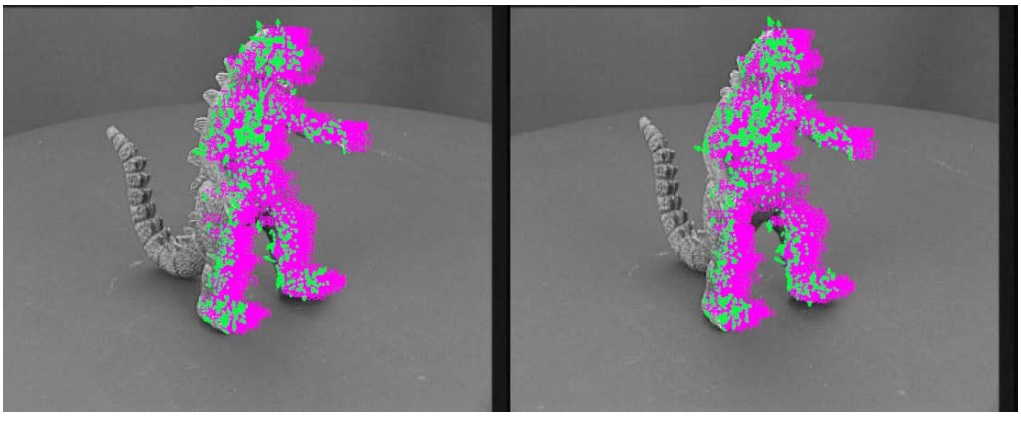
We calculate the geometric error using the optimal triangulation, as illustrated in **Fig. 2**. For accuracy, we only use points that are smaller than the value of 0.5 pixels. This is because SIFT has an accuracy of half a pixel.

The features obtained through this method satisfy the epipolar consistency. To filter out these matches, the retained matches must be compatible with the epipolar geometry.

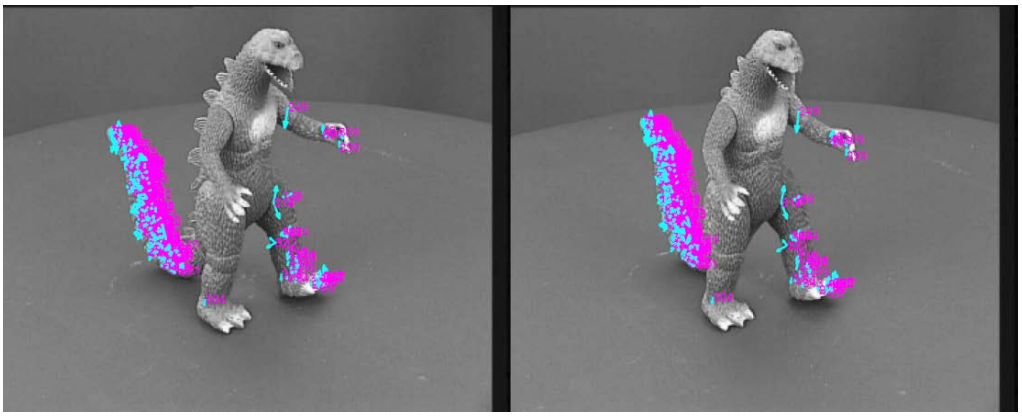
2.4.3 Clustering

Depending on the circumstances, diverse homographies can be found in a single image. Thus, we require a sufficient and appropriate distribution for the areas we seek to reconstruct. However, as indicated in **Fig. 3(a)**, the tail area cannot be matched as the dinosaur changes its pose. Although there are many matching points in the body area, as can be seen in the next section, the lack of geometric data in the tail area results in an abnormal reconstruction.

To solve this problem, we locate the matching points by repeating the three steps in Section 2.4.2. We determine the matching points using the points excluding the matching points found from the first result. Then, as indicated in **Fig. 3(b)**, the majority of the parts are reconstructed after the second matching. However, not all images can be found after two attempts, and four attempts are therefore typically recommended. This approach can be used to enhance the regional accuracy in incremental SFM.



(a) Matching result of first clustering



(b) Matching result of second clustering

Fig. 3. Example of multiple local homography

2.5 SFM

SFM has an advantage of a direct dense reconstruction by matching virtually all parts of the object. The tracking matrix $\mathbf{W}$ can be generated from the matching points of the images. As in SFM, $\mathbf{W}$ can be decomposed through CSF using (8).

$$\mathbf{W} = \mathbf{MS} \tag{8}$$

In [11], unlike previous methods that compute both $\mathbf{M}$ and $\mathbf{S}$ from the tracking matrix $\mathbf{W}$, $\mathbf{S}$ (structure) is then computed once $\mathbf{M}$ (motion) has been estimated based on the Levenberg–Marquardt-Subspace (LM-S) algorithm.

$$\mathbf{S} = \mathbf{M}^{\dagger}\mathbf{W} \tag{9}$$

where † denotes the Moore–Penrose pseudo-inverse.

Algorithm 1. Levenberg–Marquardt-Subspace (LM-S)

1: $\mathbf{M} \leftarrow$ **initial matrix** $(\mathbf{M_0})$

2: $\boldsymbol{\delta} \leftarrow$ **initial damping scalar** $(\boldsymbol{\delta_0})$

3: **repeat**

4:

compute gradient $(\mathbf{g})$ and Hessian (H) **from Jacobian terms** $(J_j=[s_j^T\ 1]\otimes P_j^{\perp}\Pi_j)$

5: **repeat**

6: $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} \times \mathbf{10}$

7: **find** $\Delta\mathbf{M}$ **from** $\boldsymbol{vec}(\Delta\mathbf{M}) \leftarrow (\mathrm{H} + \boldsymbol{\delta}\mathbf{I})^{-1}\mathrm{g}$

8: **until** $\boldsymbol{f}(\mathbf{M} - \Delta\mathbf{M}) < f(\mathrm{M})$

9: $\mathbf{M} \leftarrow \mathbf{M} - \Delta\mathbf{M}$

10: $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} \times \mathbf{10^{-2}}$

11: **orthogonalize M (keep mean vector t unchanged)**

12: **until convergence**

According to Gotardo *et al.* [11], it is preferable to use a predefined basis (B) for an efficient estimation. Here, B uses a discrete cosine transform (DCT) and changes as illustrated in (10).

$$\mathrm{M} = \mathrm{BX} \tag{10}$$

Because we desire a direct Euclidean (coarse-to-fine) reconstruction, we use the weak-perspective camera model. Moreover, W includes missing data, and each row displays the smooth time-trajectory of the two-dimensional (2D) points. The Jacobian term in Algorithm 1 changes as indicated in (11).

$$\mathrm{J_j} = ([\mathrm{s_j^T}\ 1] \otimes \mathrm{P_j^{\perp}} \Pi_\mathrm{j})\widetilde{\mathrm{B}}_{\mathrm{wp}} \text{ when } \mathrm{B} = \widetilde{\mathrm{B}}_{\mathrm{wp}} \tag{11}$$

This is the CSF-$\mathrm{B_{wp}}$ method used for the performance test of the proposed algorithm. Finally, our quasi-dense reconstruction can be described as indicated in Algorithm 2.

Algorithm 2. Proposed method

Input:

Image pairs or video sequences

Output:

Quasi-dense reconstruction

1: Choose a point detection algorithm from affine-SIFT (ASIFT), affine-SURF (ASURF), and combined ASIFT + ASURF. The keypoints are extracted with an adaptive threshold.

2: Calculate sufficient matching points that satisfy the epipolar geometry.

3. Apply clustering such that the matching points can cover the area of interest.

4. Create tracking matrix W from the matching points and conduct a 3D reconstruction using the CSF-$\mathrm{B_{wp}}$ method.

3. Results

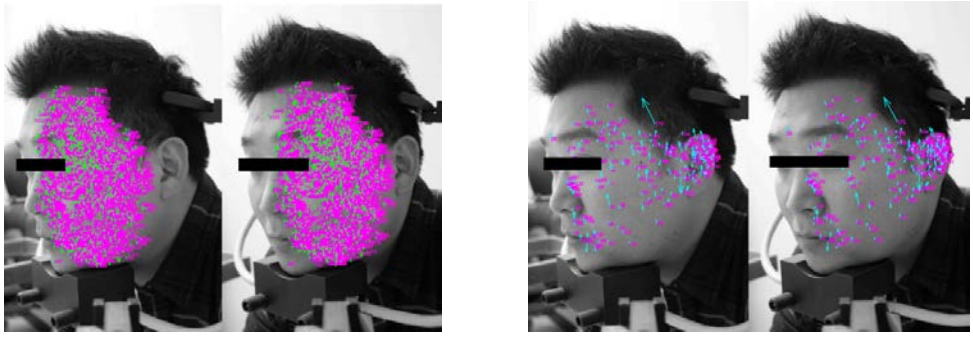
Our objective is to determine numerous and accurate matching points. We can achieve a more reliable result by eliminating outliers using multiple matching procedures. To achieve this, sufficient feature points, regardless of image type, must be available.

3.1 Face dataset

Table 2 shows the results based on the feature detector selection and affine model using face data (ten images). SIFT shows a relatively superior performance compared to SURF. Although the image is at low resolution, the Aff5-CSS results indicate a large number of points, i.e., 22,460.

Table 2. Results of CSF-Bwp with different point extraction algorithms on the face (292×438 pixels) dataset

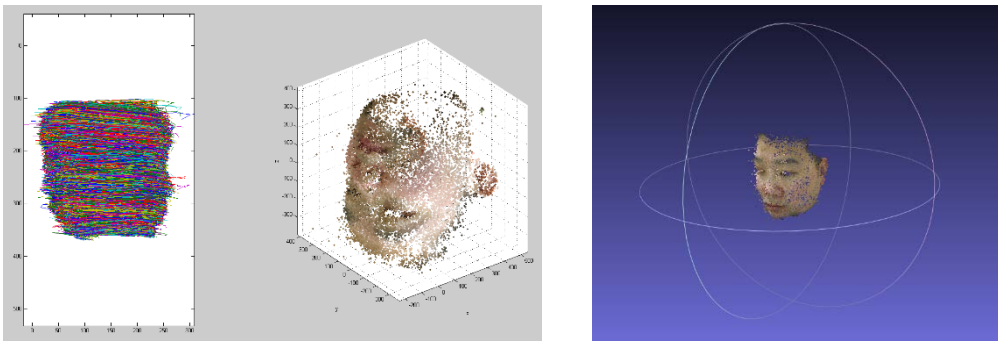
Method / Model	SIFT		SURF		Combined SIFT&SURF	
	mean /std	#pts /#%missing /RMSE	mean /std	#pts /#%missing /RMSE	mean /std	#pts /#%missing /RMSE
Original	0.1487 /0.11697	243pts /77.0% /0.411656	0.1406 /0.14096	82pts /78.8% /0.440227	0.16553 /0.12253	327pts /77.5% /0.468559
Aff1	0.12093 /0.10323	1394pts /78.4% /0.284894	0.19002 /0.12712	575pts /78.4% /0.310526	0.14468 /0.11649	2021pts /78.4% /0.320202
Aff2	0.1114 /0.095958	2910pts /78.6% /0.279037	0.17935 /0.12732	1790pts /78.3% /0.433069	0.13591 /0.11326	4797pts /78.5% /0.299193
Aff3	0.11217 /0.098766	4828pts /78.8% /0.270047	0.17275 /0.12879	4524pts /78.3% /0.355119	0.14266 /0.11653	9201pts /78.6% /0.313637
Aff4	0.11083 /0.096608	6170pts /78.9% /0.269779	0.17581 /0.12895	8519pts /78.6% /0.340973	0.15071 /0.12097	15084pts /78.7% 0.314412
Aff5	0.10895 /0.095509	6971pts /79.0% /0.261413	0.17349 /0.12904	15907pts /78.8% /0.343893	0.15964 /0.12424	22460pts /78.9% 0.326755



(a) Corresponding points of first clustering (b) Corresponding points of second clustering

Fig. 4. Result of a face image pair

Fig. 4 shows the matching results of one of the pairs in the face images. As the results indicate, **Fig. 4(a)** covers the majority of the side-profile and **Fig. 4(b)** covers most of the ear, indicating dense matching for the majority of the area of interest. **Fig. 5** shows the result of CSF-B_{wp} using data obtained from Aff5-CSS.



(a) Reconstructed 2D point trajectories W **(b) 3D reconstruction**

Fig. 5. Results of the face sequences

Our algorithm can produce excellent results for images on private websites on the Internet, most of which are low resolution. It can also be a solution to the low resolution of endoscopic images, specifically capsule endoscopy.

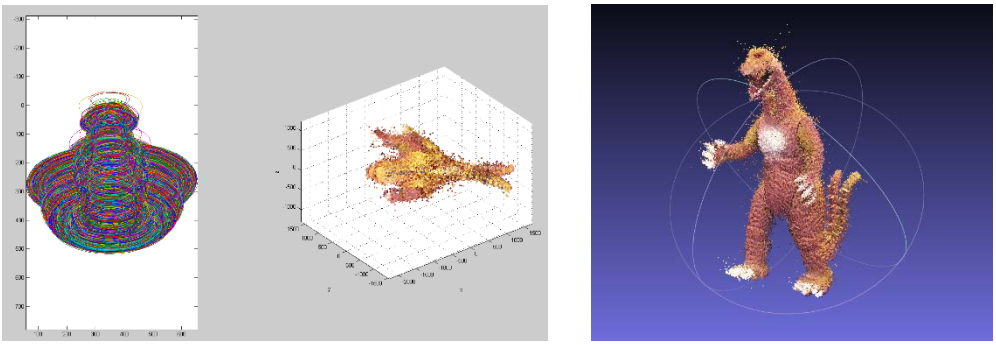
3.2 Dinosaur dataset

To evaluate the performance level of the proposed algorithm, we compared the results with those of Gotardo *et al.* [11]. The image used is that of a dinosaur. The image rotates 360° and consists of 36 images in 10° intervals. The first three values in **Table 3** comprise a dataset obtained using the Kanade–Lucas–Tomasi (KLT) feature tracker. They have 319, 2683, and 4983 points for a sparse reconstruction and an root mean square error (RMSE) value of greater than 1. By contrast, we have 213,568 sufficient points for a dense reconstruction and an extremely small RMSE of approximately 15% of the comparative error value.

Table 3. Results of CFS-Bwp on different subsets of the dinosaur (720×576 pixels) dataset

Data	319 point tracks (76.9% missing)	2,683 point tracks (87.8% missing)	4,983 point tracks (90.8% missing)	213,568 point tracks (93.8% missing)	262,409 point tracks (93.8% missing)
RMSE	1.3031	1.4833	1.2641	0.187835	0.268765

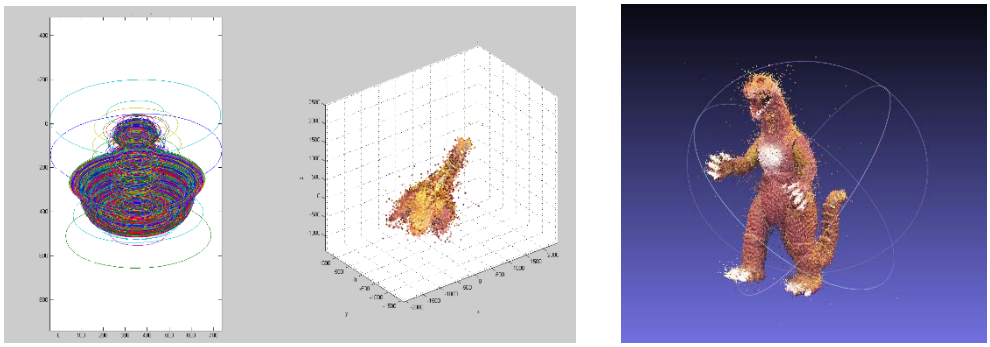
Fig. 6 illustrates why the clustering mentioned in Section 2.4.3 is necessary. As indicated in **Fig. 3**, we were able to conduct a dense reconstruction using a dataset that did not determine further matching points from the second clustering matching; however, the tail area is separated into two parts.



(a) Reconstructed 2D point trajectories W **(b) 3D reconstruction**

Fig. 6. Results of the dinosaur dataset without clustering

As indicated in **Table 3**, the RMSE value is the smallest. Although, it is preferable to have a smaller error, not all morphological aspects are represented. Several homographies are possible, and therefore a clustering method is required. The last dataset is made up of the final results, and shows a slight increase in the RMSE of 0.268765. However, an excellent dense reconstruction is achieved, as indicated in **Fig. 7**. Our results show that the proposed approach is significantly superior to other state-of-the-art algorithms.



(a) Reconstructed 2D point trajectories W **(b) 3D reconstruction**

Fig. 7. Results of the dinosaur dataset with clustering

4. Conclusion

By providing reliable geometric constrained quasi-dense matching points, the proposed algorithm provides an accurate quasi-dense reconstruction in application areas where CSF-B_{wp} can be applied. In particular, it can be employed in security systems because an iterative closest point method applied to 3D face reconstruction allows for more accurate face recognition. Because the algorithm can also be applied to low-resolution images, it has a wide range of uses. VisualSFM [20] is a representative software with a high level performance; however, it requires high-resolution images. The proposed algorithm is flexible in terms of resolution and can be applied to various imaging systems.

For improved results, apparent outliers can be determined during the 3D reconstruction. Further studies should be conducted to determine how such points can to eliminated. If we can segment a silhouette or an area of interest, eliminating the back-projection could be one such method. An excessive number of calculations also requires an improvement in speed by applying the parallel computing capability of a GPU.

In future research, the proposed method will be applied using incremental SFM, which is less restricted to a 3D reconstruction than the method using CSF and has a higher utilization. In addition, we will conduct a comparative study on the matching results by applying a deep learning approach.

References

[1] B. Shin and J. Seo, "Experimental Optimal Choice Of Initial Candidate Inliers Of The Feature Pairs With Well-Ordering Property For The Sample Consensus Method In The Stitching Of Drone-based Aerial Images," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 4, pp. 1648-1672, 2020. Article (CrossRef Link)

[2] Y. Rao, X. Ding and B. Fan, "An Efficent Method of Binocular Data Reconstruction," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 9, pp. 3721-3737, 2015. Article (CrossRef Link)

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91-110, 2004. Article (CrossRef Link)

[4] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, "Speeded-up robust features (SURF)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346-359, 2008. Article (CrossRef Link)

[5] J. Zhu, W. Sun, B. Guo and C. Li, "Surf points based Moving Target Detection and Long-term Tracking in Aerial Videos," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 11, pp. 5624-5638, 2016. Article (CrossRef Link)

[6] Z. Zhang and W. S. Lee, "Deep Graphical Feature Learning for the Feature Matching Problem," in *Proc. of International Conference on Computer Vision (ICCV)*, pp. 5086-5095, 2019. Article (CrossRef Link)

[7] N. Ufer and B. Ommer, "Deep Semantic Feature Matching," in *Proc. of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5929-5938, 2017. Article (CrossRef Link)

[8] K. Choi, J. Oh, S. Choi, and M. Kim, "A robust Human Face Recognition algorithm for flexible situations using SIFT," in *Proc. of International Forum on Medical Imaging in Asia*, 2009.

[9] J. Oh, H. Kim, J. Koo, J. Yu, T. Kang, J. Lee, and M. Kim, "ROBPCA-SIFT: a feature point extraction method for the consistent with epipolar geometry in endoscopic images," *Image and Vision Computing New Zealand New Zealand*, 2006.

[10] J. Oh, H. Kim, S. Choi, K. Choi, S. Ha, O. Lee, and M. Kim, "A robust method of feature extraction from noised endoscopic images," in *Proc. of International Forum on Medical Imaging in Asia*, 2009.

[11] P. F. Gotardo, and A. M. Martinez, "Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 2051-2065, 2011. Article (CrossRef Link)

[12] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Generic and real-time structure from motion using local bundle adjustment," *Image and Vision Computing*, vol. 27, no. 8, pp. 1178-1193, 2009. Article (CrossRef Link)

[13] A. Irschara, C. Zach, M. Klopschitz, and H. Bischof, "Large-scale, dense city reconstruction from user-contributed photos," *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 2-15, 2012. Article (CrossRef Link)

[14] M. Lhuillier, and S. Yu, "Manifold surface reconstruction of an environment from sparse Structure-from-Motion data," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1628-1644, 2013. Article (CrossRef Link)
[15] Y. Furukawa, and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362-1376, 2010. Article (CrossRef Link)
[16] M. I. Lourakis, and A. A. Argyros, "SBA: A software package for generic sparse bundle adjustment," *ACM Transactions on Mathematical Software (TOMS)*, vol. 36, no. 1, pp. 2, 2009. Article (CrossRef Link)
[17] J.-M. Morel, and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438-469, 2009. Article (CrossRef Link)
[18] Y. Pang, W. Li, Y. Yuan, and J. Pan, "Fully affine invariant SURF for image matching," *Neurocomputing*, vol. 85, pp. 6-10, 2012. Article (CrossRef Link)
[19] R. Hartley, and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2004. Article (CrossRef Link)
[20] C. Wu, "VisualSFM: A visual structure from motion system," *http://ccwu.me/vsfm/*, 2013.

Jangseok Oh received the B.S., M.S., and Ph.D. degrees in Electronics & Information Engineering from Korea University, Korea in 2004, 2006, and 2016 respectively. From Sept. 2016 to Mar. 2018, he worked at Center for Robotics Research in Korea Institute of Science and Technology(KIST). He currently is working in Agriculture Robotics & Automation Research Center, Interactive Robotics R&D Division, Korea Institute of Robotics & Technology Convergence(KIRO). His main research interests include 3D reconstruction, image processing, navigation, deep learning, and agriculture robotics.

Hyunggil Hong received the B.S., M.S., and Ph.D. degrees in Division of Electronics & Electrical Engineering from Dongguk University, Dongguk in 2012, 2014, and 2018 respectively. He currently is working in Agriculture Robotics & Automation Research Center, Interactive Robotics R&D Division, Korea Institute of Robotics & Technology Convergence(KIRO). His main research interests include image processing, deep learning, autonomous driving, and agriculture robotics.

Yongjun Cho received the B. S., M. S. in Bio Electronic Engineering form Andong National Unversity, Korea, in 2007 and 2009 respectively. He is currently a Ph.D candidate of department of Bio-ICT engineering from Andong National Unversity, Korea. He is currently working in Agriculture Robotics & Automation Research Center, Interactive Robotics R&D Division, Korea Institute of Robotics & Technology Convergence(KIRO). His main research interests include curcuit design, sensor, system intergration and robotics.

Haeyong Yun received the B.S., M.S., and Ph.D. degrees in Department of Mechanical system Engineering, Andong National University, Korea in 2009, 2012, and 2017 respectively. He currently is working in Interactive Robotics R&D Division, Korea Institute of Robotics & Technology Convergence(KIRO). His main research interests include Gantry robot control, additive manufacturing, experimental analysis, and agriculture robotics.

Kap-Ho Seo received a B.S. degree in Electrical Engineering from Korea University, Seoul, Korea, in 1999, and M.S., Ph. D. degree in Electrical and Electronic Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2001 and 2009, respectively. He is currently working as a chief researcher at KIRO since 2009. Also, He has been working as an adjunct professor at the Department of Mechanical Engineering at Pohang University of Science and Technology (POSTECH), Pohang, Korea since 2020. His fields of interest are intelligent control, embedded systems, image processing, and robotics.

Hochul Kim received his B.S. degree in Medical Electronics and M.S. degree in Biomedical Engineering from Korea University, Korea, in 2002, 2004, respectively. He received the Ph.D. degree in Biomedical Engineering from Korea University in 2009. Since 2012, he has been with the Department of Radiological Science, Eulji University, Republic of Korea, where he is currently a professor. His research interests include Medical Image Processing, Radiation Detection, and Neural Network.

Mingi Kim graduated from the Department of Electrical Engineering at Korea University B.S and received a M.S. in Communication Theory at University of Columbia, and with a Ph.D. in Image Processing from Polytechnic University in 1991.

Onseok Lee graduated from Korea University with a B.S., M.S., and Ph.D. in Biomedical Engineering in 2011, where he developed biomedical engineering techniques such as computer vision, haptics, and molecular imaging