Feature vector for multiple perspective independent point cloud registration

Chengsi Zhang
Electrical and Computer Engineering
Faculty of Engineering and Applied Science
Memorial University
cz2075@mun.ca

Stephen Czarnuch
Electrical and Computer Engineering
Faculty of Engineering and Applied Science
Memorial University
sczarnuch@mun.ca

***Abstract*—We propose a novel feature vector towards perspective independent scene registration using 3D sensors. We captured both RGB and depth data from multiple 3D sensors placed arbitrarily, leading to perspectives that were not known to our system. Our previous work on perspective independent ground plane estimation provided the foundation for our current work. Our perspective independent ground plane estimation algorithm produces 3D point clusters representing all moving persons, and all static objects in the room. Based on the unique geometric information relating moving objects to static objects, we define a global 3D feature to represents the objects' spatial relationships within a 3D point cloud. We then use this novel vector to determine the rotation around the y-axis and translation along the x- and z- axes, revising the pre-alignment we first completed using the ground plane. After using this feature in our registration algorithm, we successfully find the estimated transformation matrix between two point clouds for 13 out of 16 data trails.**

***Index Terms*—Human identification, homography, homography decomposition, point cloud segmentation, 3D data analysis.**

I. Introduction

At present, three dimensional (3D) data has been utilized for many applications including interactive modeling, virtual display, and digital reconstruction, which have widespread utilization in many industries including training and simulation[1][2], augmented reality (AR)/virtual reality (VR)[3][4], gaming[5] or movies[6]. However, because 3D data provide more information, they also requires more processing time and memory storage compared to traditional two dimensional (2D) data. In addition, due to the limitations of field of view (FOV), accuracy requirements, and coverage requirements, a full 3D digital copy of a scene or object usually consists of multiple partial 3D data frames that are combined by 3D registration algorithms.

3D registration is the process of finding the possible transformation matrix between two 3D data frames, so that the common data in these two 3D data frames can be aligned together. This process essentially involves finding the feature descriptors in two 3D data frames, matching the feature descriptors in one frame to another based on similarity, and finally calculating the rotation, translation and scaling when similar feature descriptors align together. Existing feature descriptors can be broadly separated into two categories: local feature descriptors and global feature descriptors.

A. Local feature descriptors

Local approaches first try to identify the key points that may represent the most valuable information, and then construct a group unique identifiers from the 3D data frame, where each unique identifier is constructed from a small cluster of points associated with key points[7]. After determining the feature descriptors, the transformation matrix is calculated by estimating the best solution for all the descriptor matches. Since the local feature descriptors are generated from the key points, the accuracy of finding correct key point largely decides the performance of registration results based on local feature descriptors. Local descriptors share some common issues including the balance between low descriptiveness and high computational complexity, and high sensitivity to the noise and distortion in the 3D data frame[9].

B. Global feature descriptors

Unlike the local feature descriptor that is based on small clusters of points, the global feature descriptor summarizes the useful geometric information on the entire data frame level. After constructing the global feature descriptor for two 3D data frames, an optimal transformation solution is calculated across the whole point cloud by minimizing the distance between two global feature descriptors. Due to the global feature descriptor usually relying on segmentation algorithms (whose performance is highly sensitive to errors and noise) to extract the most representative object information from the 3D data frame[8], local non-linear refinement is often applied after the registration based on the global feature descriptor. Compared to the local feature descriptor, the global feature descriptor has higher tolerance for larger values of initial position difference between two 3D frame. However, since this type of descriptor is generated from the selection of partial data that might be useful, this descriptor is highly likely to result in incorrect or insufficient correspondences[10].

In this paper, we propose a novel global feature vector targeting the rigid transformation in perspective independent 3D registration, which uses the output information of our previous ground plane detection algorithm[11]. Our ground plane detection algorithm estimates the position of the ground plane in each sensor independently, as well as segmenting any moving people and other static objects in the indoor environment. Based on these data, our new global feature is able to register two 3D data frames captured from perspective independent cameras in a cluttered, dynamic and complex indoor environment if minimum assumptions are met:

- At least one person can be seen smoothly moving in both 3D sensors' FOVs
- The moving person's body is perpendicular to the ground
- The perspective and position of both cameras remain unchanged.

II. Related work

Both local and global feature descriptors have numerous existing implementations. One of the most popular local feature descriptors is Fast Point Feature Histograms (FPFH)[12]. A simplified point feature histogram (SPFH) for each point in the 3D data frame is first generated by constructing three histograms from each point and its neighbour points along three dimensions, then FPFH is built based on weighted sum of the SPFH of a feature point and the SPFHs of the points in the feature point's support region. The dimension of FPFH is $3d$ where d is the number of bins along one dimension[12]. Local Surface Path(LSP) is one of the fundamental local feature descriptors. Two key factors of LSP are the shape index of each point p in the support region of a key point k and the cosine value of the angle between the surface normal of the point q and the normal at the key point k. The LSP descriptor is calculated by accumulating points in particular bins along the two dimensional coordinates formed by these two key factors. Another useful local feature descriptor that has widespread utilization is Signature of Histogram of Orientations (SHOT)[13] which is based on Local Reference Frame(LRF)[13]. The LRF is constructed for each key point k and its neighbour points which are in its support region. After dividing the support region into three volumes along radial, azimuth and elevation axes, a local histogram is calculated by accumulating points counts into bins based on the angles between the normal of neighbour's points and the normal of the key point k in each volume. Finally, the SHOT descriptor is generated by combining all the local histograms. Signature of Histogram of Orientations for Color (SHOT COLOR) is then extended based on SHOT approach to work with texture data[14].

Similar to the local feature descriptors, the global feature descriptor also has numerous implementations. Marianna and et al[15] encodes the 3D data frame by using Global Structure Histograms (GSH). This global feature starts with constructing local descriptors based on surface-shape characteristics of 3D data, and then labelling the surface class for each point by using k-means algorithm followed by the computation of Bag-of-Words model. After the relationship between different classes is determined by triangulation, GSH can be formed as histogram based on the distribution of distance along the surface. One of the recent global feature descriptors built based on SHOT is Scale Invariant Point Feature(SIPF) [16] which represents the object or scene with border shape encoding. It first encodes the object border by combining LRF and covariance matrix which is defined in the SHOT[13], then it computes the feature value $q^* = argmin_q\|p - q\|$ between feature point p and the edge point q as the reference direction. After dividing the angle q^* of local cylindrical coordinates into N regions, whose angle is within $2\pi iN$ and $2\pi(i+1)N$ for i = 0, 1, ..., N-1, SIPF descriptor is constructed by concatenating all the normalized cell features $D_i = exp(\frac{d_i}{1-d})$, where d_i is the minimum distance between a point p and the ith region. Another recent global feature descriptor with promising computational time and robustness to Gaussian noise is Global Orthographic Object Descriptor (GOOD)[17]. GOOD first defines a unique but repeatable LRF based on the Principle Component Analysis, and then the data in the 3D frame is orthographically projected onto three planes that are constructed as the X-Z, X-Y and Y-Z axes respectively. Each plane is divided into multiple bins and a distribution matrix is computed by counting the number of points for each bin. Finally, the GOOD descriptor is constructed by concatenating the entropy and variance vectors from each distribution matrix across the whole 3D frame.

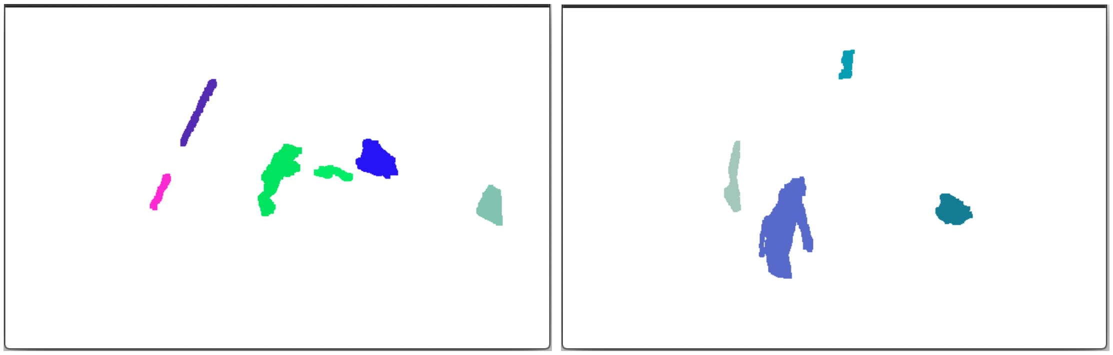
However, in the scenario our application is targeting, which is in clustered, dynamic and complex indoor environments, none of the existing local or global features can properly register the 3D data frames from two perspective independent sensors. Therefore, with the inspiration from [18], we propose a novel global feature that describes the uniqueness of an indoor geometry feature and utilizes the ground plane as the reference plane.

III. Methodology

Since our new global feature descriptor is built based on the output data that is generated from our previous estimating ground plane work and we also use the ground plane as the reference plane, this section first briefly describes our previous work.

A. Ground plane estimation

Our ground plane estimating algorithm aims to identify and segment the ground plane in a scene where the camera has an arbitrary and unknown orientation and location, with the assumption that at least oen person can be seem smoothly moving in the camera FOV. The whole process can be separated into three steps: 3D data segmentation, moving object identification, and ground plane estimation. During 3D data segmentation, all large planes are extracted as possible ground plane candidates P_i and all isolated point clusters are segmented as objects O_i in the indoor environment after all the large planes are removed. We then use Scale-invariant feature transform (SIFT) based Motion-Split-And-Merge (MSAM)[19] to find the moving person(s) among the object group O_i and the approximate moving trajectory vector $\vec{t}$ of each moving person within a block of frames B_i. Finally, we estimate the ground plane based on a cascaded filter



(a) left camera objects

(b) right camera objects

Fig. 1: 3D data captured from two sensors

which mainly relies on the geometric relationship between the ground plane and other objects. Therefore, the output of the ground plane estimation algorithm consists of the most likely ground plane GP, the moving human(s) M_i, the static objects O_j, and the estimated trajectory vector for each moving human $\vec{t}_i^k$ in the kth block of frames.

B. Global feature descriptor

After we apply the ground plane estimation algorithm to two 3D data frames from two perspective independent sensors, we obtain information about the ground plane, moving human(s) and a trajectory vector of each moving human within a block of frames, and static objects (e.g., see Figure 1). With the assumption a moving human walks through an area common to the two camera's FOVs, we can align the two 3D data frames based on the ground plane, or more specifically by using the y-axis translation, z-axis rotation (yaw) and x-axis rotation (roll) of the ground plane. Hence, by using first aligning the two frames to their ground planes, the entire 3D transformation problem is simplified to the 2D problem of finding the remaining x-axis and z-axis translations and a y-axis rotation (pitch). By converting the 3D registration challenge to a 2D problem, all of the vectors in 3D world we use to construct our global feature descriptors are reduced to 2D vectors by removing the y-axis value from the original three dimensional vector. In other words, we can now describe the two dimensional vectors as if we are viewing the 3D data frame from an overhead perspective. From this perspective, our vector is then simply the set of human-to-object vectors within the scene.

Depending on the accuracy of 3D data segmentation and the organization of the objects in the room, objects in the indoor environment are separated into multiple point clusters. Therefore, the first step for constructing our global feature is building the human-to-object vectors $M\vec{O}_{ij}$ that point from the center of mass M_i^{center} of each moving human M_i to the center of mass O_j^{center} of each static object O_j, as shown in Figure 2. This creates a 2D vector pool containing $i \times j$ vectors of geometric description, where i is the number of moving humans and j is the number of objects. Within the set of all object point clusters O_i, multiple point clusters may actually represent the same real object (e.g., be incorrectly separated by segmentation). To overcome this, if objects are close to each other *and* have similar euclidean distance to the human, we merge the two human-to-object vectors $M\vec{O}_{ij}$ and $M\vec{O}_{ik}$ by building a new vector $M\vec{O}_ijk$ whose origin is M_i^{center} and the center of mass point $O_j^{center}k$ of two point clusters O_j and O_k, if $\|M\vec{O}_{ij} - M\vec{O}_{ik}\| < \varepsilon$ and $\angle(M\vec{O}_{ij}, M\vec{O}_{ik}) < \alpha$, where ε and α are the euclidean distance threshold and angle threshold respectively.

After combining the redundant vectors from the 2D vector pool, we construct a global feature matrix F_i for the ith moving person by combining all the human-to-object vectors $M\vec{O}_i$ that start from the center of mass point M_i^{center} of the moving human M_i at time t.

$$F_i = \begin{bmatrix} M\vec{O}_i0 \\ \cdots \\ M\vec{O}_in \end{bmatrix}, n = 0, 1, \cdots, J \tag{1}$$

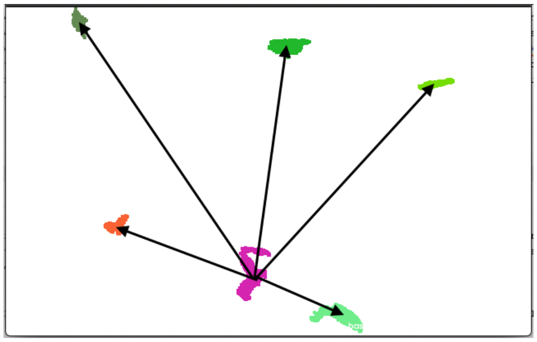


Fig. 2: Human-to-object vectors from top view

Since the approximate trajectory vector for each human within a block of frames is known, we determine whether the moving human M_i in the first camera's frame is possibly the same moving human M_i' in the second camera's frame by checking the number of row pairs (between the global feature matrix F_i and F_i') that have a mean squared error (MSE) that is lower than the threshold μ. We achieve this by calculating the mean squared error between all the possible pairs between a vector in F_i and all existing vectors in F_i'. If the mean square error between a vector $\vec{MO}_{ij}$ in F_i and another vector $\vec{MO}'_{nm}$ in F_i' if above μ, we remove these two vectors from the global feature matrix. By comparing all possible combinations, we conclude that the global feature matrix represents the same person in both cameras if the number of vector pairs with an MSE below μ is larger than the threshold ν. We use the known approximate trajectory vector as the reference direction (i.e., the trajectory vector should have the same direction for the same person), and we find the y-axis rotation (pitch) matrix by calculating the average rotation angle after aligning all the vector pairs with an MSE below μ. Finally, we calculate the x-axis and z-axis translations by matching the point cluster of the moving human M_i in the two camera frames.

Similar to other global feature descriptors, the transformation matrix that our feature descriptor generates usually contains noise and error. Therefore, the final step of our global feature descriptor registration is the iterative closest point (ICP) [20] algorithm for refining the transformation of the roughly aligned 3D data frames.

IV. Discussion and future work

In this paper, we proposed a novel global feature descriptor for registering multiple 3D data frames captured from multiple perspective independent 3D sensors, only assuming that at least one moving person can be seen in the common FOV of the sensors, and that the human's body is perpendicular to the ground plane while moving. Our global feature descriptor is targeted toward indoor scenes that are highly cluttered, dynamic and complex, where other existing local and global feature descriptors do not provide good result based on their limitations. One of our most significant limitations is that due to the process of constructing our global feature, our approach will not work in the scenarios where only a human and ground plane appear in the common FOV of 3D sensors. In the future, we will focus on improving the efficiency and complicity of this global feature generation; applying our global feature to more real datasets; verifying the success rate of our global feature against common sources of noise; and comparing the computation time and accuracy with other local and global features.

References

[1] Waran, Vicknes, et al. ”Injecting realism in surgical training—initial simulation experience with custom 3D models.” Journal of surgical education 71.2 (2014): 193-197.

[2] Bruzzone, Agostino G., and Francesco Longo. ”3D simulation as training tool in container terminals: The TRAINPORTS simulator.” Journal of Manufacturing Systems 32.1 (2013): 85-98.

[3] Garon, Mathieu, et al. ”Real-time high resolution 3D data on the HoloLens.” 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct). IEEE, 2016.

[4] Cox, Donna J., Robert M. Patterson Jr, and Marcus L. Thiebaux. ”Virtual reality 3D interface system for data creation, viewing and editing.” U.S. Patent No. 6,154,723. 28 Nov. 2000.

[5] Adams, Khaled. ”3D enhancements to gaming components in gaming systems with real-world physics.” U.S. Patent No. 10,115,261. 30 Oct. 2018.

[6] Emoto, Michiko, Shinya Miyamoto, and Kazuyoshi Yamamoto. ”Navigation apparatuses, methods, and programs for generation of a 3D movie.” U.S. Patent No. 7,974,781. 5 Jul. 2011.

[7] Guo, Yulan, et al. ”A comprehensive performance evaluation of 3D local feature descriptors.” International Journal of Computer Vision 116.1 (2016): 66-89.

[8] Chen, Jingdao, Yihai Fang, and Yong K. Cho. ”Performance evaluation of 3D descriptors for object recognition in construction applications.” Automation in Construction 86 (2018): 44-52.

[9] Yang, Jiaqi, Zhiguo Cao, and Qian Zhang. ”A fast and robust local descriptor for 3D point cloud registration.” Information Sciences 346 (2016): 163-179.

[10] Gelfand, Natasha, et al. ”Robust global registration.” Symposium on geometry processing. Vol. 2. No. 3. 2005.

[11] Zhang, Chengsi, and Stephen Czarnuch. ”Perspective Independent Ground Plane Estimation by 2D and 3D Data Analysis.” IEEE Access 8 (2020): 82024-82034.

[12] Rusu, Radu Bogdan, Nico Blodow, and Michael Beetz. ”Fast point feature histograms (FPFH) for 3D registration.” 2009 IEEE international conference on robotics and automation. IEEE, 2009.

[13] Tombari, Federico, Samuele Salti, and Luigi Di Stefano. ”Unique signatures of histograms for local surface description.” European conference on computer vision. Springer, Berlin, Heidelberg, 2010.

[14] Salti, Samuele, Federico Tombari, and Luigi Di Stefano. ”SHOT: Unique signatures of histograms for surface and texture description.” Computer Vision and Image Understanding 125 (2014): 251-264.

[15] Madry, Marianna, et al. ”Improving generalization for 3d object categorization with global structure histograms.” 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012.

[16] Lin, Baowei, et al. ”Scale invariant point feature (SIPF) for 3D point clouds and 3D multi-scale object detection.” Neural Computing and Applications 29.5 (2018): 1209-1224.

[17] Kasaei, S. Hamidreza, et al. ”GOOD: A global orthographic object descriptor for 3D object recognition and manipulation.” Pattern Recognition Letters 83 (2016): 312-320.

[18] Taguchi, Yuichi, et al. ”Method for registering points and planes of 3D data in multiple coordinate systems.” U.S. Patent No. 9,183,631. 10 Nov. 2015.

[19] Dragon, Ralf, Bodo Rosenhahn, and Jörn Ostermann. ”Multi-scale clustering of frame-to-frame correspondences for motion segmentation.” European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.

[20] Besl, Paul J., and Neil D. McKay. ”Method for registration of 3-D shapes.” Sensor fusion IV: control paradigms and data structures. Vol. 1611. International Society for Optics and Photonics, 1992.