HIGH-SPEED STRUCTURED LIGHT SCANNING SYSTEM AND 3D GESTURAL POINT CLOUD RECOGNITION

Yin Zhou¹, Kai Liu¹, Jinglun Gao¹, Kenneth E. Barner¹, Fouad Kiamilev¹

¹University of Delaware, Newark, DE, USA, 19716
²School of Electrical Engineering and Information, Sichuan University, China

ABSTRACT
In computer-vision-based human computer interaction (HCI), higher-quality signal leads to better system performance. In this paper, we develop a real-time high-resolution 3D object scanning system based on structured light illumination (SLI). Our system fuses depth information with RGB texture to reconstruct high-resolution 3D point cloud. The point cloud preserves accurate surface geometry of the object (e.g., finger postures of hands, facial expressions, etc). Respectively, for a 640 × 480 video stream, our system can generate phase and texture video at 1500 frames per second (fps) and produce full 3D point clouds at 300 fps. For gesture recognition, we propose to combine the module of robust face recognition with the module of 3D point cloud classification. Moreover, rather than extracting sophisticated features, we leverage the accurate reconstruction and classify each point cloud by directly matching the whole 3D surface geometry with the templates of different classes. The proposed recognition system is robust to the scaling, translation, rotation and texture of objects. Finally, utilizing the system, we contribute to the research community two large-scale high-resolution 3D point cloud databases, i.e., SLI 3D Hand Gesture Database and SLI 3D Face Database. The proposed point cloud recognition approach achieves recognition rates up to 98.0% over the gesture database and 88.2% over the face database in our pilot study.

Index Terms— HCI, Structured Light, Gesture Recognition, Face Recognition, Multi-modal, Benchmarks

1. INTRODUCTION

Automatic gesture recognition has been a widely studied area and can become the mainstream tool in the future for humans to contactlessly interact with computers, robots or virtual environment. Traditional data acquisition methods vary from cyberglove [1] and marker-based tracking system [2] to computer vision techniques using a single camera [3] and camera arrays [4], to thermal imaging technology [5]. In recent years, range camera with depth information has raised great interest in the area of 3D gesture recognition. Microsoft Kinect has provided a low-cost choice for depth sensor, which however is of limited depth resolution [6] and low processing speed in generating 3D point clouds (30fps).

In this paper, we propose a system based on structured light illumination (SLI) [7] to reconstruct 3D surface shape of the object. SLI is a technique well known for the following characteristics [7–9]: 1) High accuracy in the depth measurement [8]; Real-time performance in data acquiring and processing [7, 9]; Robustness to the texture characteristics of the objects [9]; Operational safety for both the users and the scanned objects; Affordable hardware off the shelf [8]. As will be demonstrated later, our experimental system achieves a processing rate higher than 1500 frames per second (fps) to generate phase and texture video and up to 300 fps to produce high-quality 3D point clouds using just one core of an Intel Xeon W3690 processor running at 3.46 GHz.

On the gesture point cloud recognition, we propose a classification scheme based on surface shape matching, which is free of the sophisticated feature extraction process. For improving the efficiency and accuracy of recognition, we further propose a user-specific interface combining the robust 2D facial identification module and the 3D point cloud classification module. Upon surface shape matching, the fitness between two point clouds is estimated based on the Gaussian Mixture Model (GMM) via maximizing a likelihood function [10]. Within the user-specific profile, the relationship between the query sample and a template can be effectively modeled as rigid transformation, which has the advantages of low computational complexity and robustness to outliers.

Point sets registration is an important technique in computer vision, medical imaging and automatic target recognition, etc. The goal of point sets registration is to discover the correspondences between two sets of points and then based on the correspondences to solve the transformation that maps one point set to the other. Generally, the mapping relationships between two related point sets can be categorized as rigid transformation and nonrigid transformation. Rigid transformation allows only scaling, translation, rotation or
Fig. 1: Reconstructed 3D point clouds of hand and face using the proposed structured light scanning system. (Left Column) Side view, (Middle Column) Front view, (Right Column) Bottom view.

their combination. Nonrigid transformation includes the deformation to the shape of a point cloud, such as stretching, shrinking or twisting. Some representative rigid point set registration algorithms are Iterative Closest Point (ICP) based methods [11–14], probabilistic approaches [15, 16] and spectral algorithms [17, 18], etc. Recently, Myronenko and Song proposed Coherent Point Drift (CPD) algorithm for point sets registration and achieved state-of-the-art results. CPD can yield robust estimation of the similarity between two point sets. Thus it is desired to extend its capability as a shape matching engine to the area of 3D point cloud based gesture recognition.

We contribute to the research community two large-scale high-quality 3D point cloud database, i.e., SLI 3D Hand Gesture Database and SLI 3D Face Database. Our preliminary results show that the proposed classification method achieves recognition rates up to 98.0% over the gesture database and 88.2% over the face database.

In summary, our contribution is the development of a user in/dependent 3D gestural HCI interface based on high-speed SLI [7, 9], point set matching [10] and robust face recognition [19]. To our best knowledge, no such effort is reported in literature.

2. PROPOSED APPROACH

Structured light is considered as one of the most accurate techniques to recover 3D object surfaces [8]. With precisely-preserved 3D surface shape information, point cloud classification algorithms can achieve high recognition rates. In this section, we first describe the implementation of the proposed structured light scanning system and then introduce the point cloud classification algorithm.

2.1. System Implementation

We implement the proposed system using Microsoft Visual Studio 2008 with managed C++. The system is equipped with a high-speed color camera and a projector. The imaging sensor is an 8-bits-per-pixel, Prosilica GC655C, gigabit ethernet camera with a frame rate of 90 fps and 640 × 480-pixel resolution. The projector model is Acer X1161 DLP. The camera/projector pair is synchronized at 60Hz by the VSYN signal from the nVIDIA video card. The processing unit is a Dell Precision T3500 Workstation with an Intel Xeon W3690 Hex-core CPU running at 3.46 GHZ.

During data acquisition, the system projects a sequence of time-multiplexed light patterns onto the object at 60 fps, which is the maximum allowable frame rates of the projector for eight-bits-per-pixel video. In order to suppress the effects of sensor noise and shorten data acquisition time, we employ the dual-frequency pattern scheme [7]. The system can achieve a processing rate higher than 1500 fps in generating phase and texture information. The processing unit performs real-time decoding on the captured images to extract 3D coordinate information of the scanned object surface geometry at a benchmarked rate of up to 300fps for 640 × 480-pixel video frames. The impressive 3D reconstruction frame rate is attributed to implementing the decoding equations as a series of lookup tables (LUTs) [7], which is an order of magnitude faster than traditional decoding methods based on matrix inversion1. Finally, the reconstructed 3D point clouds are displayed in a real-time fashion at the maximum 60fps frame rate of the video card. Because the frame rate of the projector is only 60Hz, the final speed of our system is limited at 60Hz. Exemplar 3D reconstructions of hand gestures and faces are demonstrated in Fig. 1. The point clouds are very dense, smooth, and high-resolution, which enables effective shape matching.

2.2. Gesture Recognition

We consider a gesture recognition problem of distinguishing the partial 3D shape of various hand postures. For the purpose of effective recognition, we propose a user-specific 3D gesture recognition interface (Fig. 2) by combining the module of 2D facial identification and the module of 3D gestural point cloud classification. Specifically, to initialize 3D gesture recognition, the user identity is to be first confirmed with a robust 2D face recognition program. Then, a user profile associated to the identify is invoked, which contains customized gestural point cloud templates consistent with the user’s operation characteristics, such as the hand shape, left or right handed habit, wore-on decorations, etc. Finally, gestural point clouds are recognized as one of the templates through an efficient pairwise matching engine [10].

1For detailed computational complexity comparisons between LUT and traditional decoding methods, see [7]
For face recognition, the algorithmic implementation is based on sparse representation based classification (SRC) [19]. Over a dictionary of representational face templates, the query image is sparsely coded via an efficient $\ell_1$ minimization solver (Feature-Sign Search [20]). Then the classification is performed by evaluating which class yields the minimum reconstruction error, as

$$r_i(y) = \|\Phi y - \Phi A \delta_i(\alpha)\|_2,$$  \hspace{1cm} (3)

where $\delta_i(\alpha) = [0, \ldots, \alpha_{i1}, \ldots, \alpha_{iN}, \ldots, 0]$. The class label of $y$ is determined as $\text{label}(y) = \arg \min_{i \in \{1, \ldots, K\}} r_i(y)$. In the proposed system, the facial identification module is employed once upon the login of a particular user, which typically takes less than one second.

In gesture recognition, several factors, $i.e.$, occlusion, illumination, scaling, translation and rotation, may adversely affect the performance of classification algorithms [24]. However, the proposed high-speed 3D reconstruction system, to a large extend is robust to the first two factors ($i.e.$, occlusion, illumination), as the high-resolution depth information helps distinguish overlapped regions while the active illumination eliminates the effects due to extreme illumination conditions. We denote $s \in \mathbb{R}$ as the scaling factor, $t \in \mathbb{R}^3$ as the translation vector, and $R \in \mathbb{R}^3$ as the rotation matrix. Since the user profile contains only point cloud templates conforming to the user’s hand shape, we employ rigid point set registration methods to estimate these three unknown factors.

Specifically, let $X \in \mathbb{R}^{3 \times N}$ be one of the template point set consisting $N$ points and $Z \in \mathbb{R}^{3 \times M}$ be the query point set containing $M$ points. According to [10], we consider points $z_m \in \mathbb{R}^3$ in $Z$ as the Gaussian Mixture Model (GMM) centroids and treat points $x \in \mathbb{R}^3$ in $X$ as data points generated by the GMM. Thus the GMM probability density function can be written as:

$$p(x) = \sum_{m=1}^{M} P(z_m)p(x|z_m),$$  \hspace{1cm} (4)

where $p(x|z_m) = \frac{1}{(2\pi \sigma^2)^{3/2}} \exp \frac{-|x - R z_m|^2}{2 \sigma^2}$, $P(z_m) = \frac{1}{M}$ and $\sigma^2$ is the isotropic covariance (for all $m = 1, \ldots, M$).

In order to estimate $s$, $t$ and $R$, we maximize the likelihood function or equivalently minimize the negative log-likelihood function as:

$$E(s, t, R, \sigma^2) = -\sum_{n=1}^{N} \log \sum_{m=1}^{M} P(z_m)p(x_n|z_m).$$  \hspace{1cm} (5)

Parameters in Eq. 5 can be solved using Expectation Maximization (EM) algorithm and we employ the toolbox [10] for fast implementation. $\sigma^2$ reflects the difficulty of aligning two point sets. Thus we treat $\sigma^2(X, Z)$ as a pari-wise similarity function of any two point sets $X$ and $Z$. The smaller $\sigma^2$ is, the higher similarity two point sets possess. The decision rule is given as:

$$\text{label}(Z) = \arg \min_i \sigma_i^2(X_1), \ldots, \sigma_i^2(X_i), \ldots, \sigma_i^2(X_C).$$  \hspace{1cm} (6)
where $\sigma_2^2(X_t) = \sigma^2(X_t, Z)$ and $C$ is the number of classes.

In our case, the advantage of using rigid point set registration methods over nonrigid registration methods is twofold: 1) model simplicity and lower computational complexity: nonrigid point set registration methods typically require the estimation of a large number of transformation parameters, while rigid transformation approaches only need to recover $s$, $t$, $R$ and the correspondence; 2) robustness: compared to rigid transformation based methods, the performance of non-rigid registration methods is more sensitive to noise, outliers and missing data points [10].

3. RESULTS

3.1. SLI 3D Hand Gesture Database

The constructed SLI 3D Hand Gesture Database contains 1040 high-quality dense point clouds of static finger spellings of 26 English letters, according to the American Sign Language (ASL) alphabet. Five volunteers participated in the data collection. Each individual performed all the 26 letters with static hand posture during the scanning. During the recording of each letter, every individual performed 2 upright postures with the difference only in spatial shifting, and performed another 6 different postures with both spatial shifting and rotation. Users were sitting at a distance of 80 - 100cm from the camera for satisfactory 3D reconstruction. To secure enough informative points captured per scanning, we restrict the rotation angle to be approximately within a 45° cone. As a proof-of-concept, the exemplar hand point clouds are demonstrated in Fig. 1 and Fig. 2.

To fully test the proposed gesture recognition method, we consider two classification scenarios: 1) employing only the 3D gesture classification module (GM); 2) employing the combination of 2D facial identification module (FM) and the GM. Classification in the first scenario is based on inter-user point clouds matching and thus is more challenging than in the second scenario. In each scenario, we compare the proposed rigid point set registration based classification scheme [10] with its nonrigid counterpart [10]. We also include the Iterative Closest Point (ICP) [11, 12] method into the comparison, which is a widely applied rigid point set registration algorithm due to its simplicity and low computational complexity. For the purpose of fair comparison, in each method, one sample per letter class is randomly selected as the template training data, while the remaining samples are for testing. Note that in the 2nd scenario, the FM yields 100% identification accuracy and thus the evaluation is essentially within the profile of the correct user.

Preliminary results are based on evaluating the three approaches over a subset of the database corresponding to the first 10 English letters (A – J). All recognition rates are reported based on our own implementation over 10 repetitions. It is observed from the first two columns in Table 1 that by taking FM into consideration, gesture classification accuracy is significantly improved compared with using GM alone, which verifies the effectiveness of the proposed module combination strategy. Moreover, in both scenarios, the proposed classification strategy achieves the highest recognition rates with the minimum time cost. The confusion matrix of the proposed method is shown in Table 2.

![Exemplar point clouds from the SLI 3D Face Database](image)

3.2. SLI 3D Face Database

The SLI 3D Face Database contains 1152 high-quality dense point clouds of 4 static facial expressions under 3 different view points and 2 different occlusion conditions. The database is established by a population of 24 volunteers com-

| Table 1: Comparison of recognition rates and time elapse over the SLI 3D Hand Gesture Database. Time elapse (second) represents only the time for classifying one query point cloud. |
|---|---|---|---|
| Proposed | GM | FM + GM | Time Elapse |
| Nonrigid [10] | 66.7% | 92.0% | 12.09s |
| ICP [11, 12] | 88.9% | 96.0% | 4.81s |

| Table 2: Confusion matrix of the proposed method based on rigid point set registration using only the GM module. |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J |
| A | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| B | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| C | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| D | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| E | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| F | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| G | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| H | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| I | 0.11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.89 | 0.00 |
| J | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
posing of undergard student, grad students, and professors. We collect data for each individual through two recording sessions with the same experimental setup. Within each session, we design two occlusion conditions, i.e., no occlusion and partial occlusion by a 3cm-wide ruler from a distance of approximately 50cm. Under each of the conditions, an individual is required to face the camera at 3 different view points, i.e., ±45° and 0° (up-front), while at each view point performing 4 kinds of static facial expressions, i.e., neutral, sad, happy, and anger. Examples from the face database are demonstrated in Fig.3. In the pilot study, we choose the frontal face point clouds with no occlusion from session 1 to form a subset containing 96 point clouds (4 point clouds for each individual). We investigate the aforementioned three methods in four scenarios, each associated with one expression. Specifically, in each scenario, the 24 point clouds corresponding to an expression are employed as templates while the remaining 72 point clouds are for testing. Experimental results are summarized in Table 3. The proposed method outperforms its competitors in all the scenarios, yielding the highest single scenario accuracy of 95.8% when using neutral face as templates and the highest average accuracy of 88.2%.

4. CONCLUSIONS

We have presented a high-speed SLI 3D object scanning system, which can generate high-resolution point clouds at a speed of 300 fps. Leveraging the accurate reconstruction, we propose a point cloud classification scheme, which needs no complicated feature extraction process. For practical 3D gesture recognition, a user-specific interface is introduced. By combining the facial identification module and the gesture classification module, the user interface can effectively diminish the diversity among users and reduce the model complexity, which leads more satisfactory recognition results. More importantly, we contribute the two large-scale 3D gesture and face databases to the research community. Our future goal includes achieving real-time point cloud classification and benchmarking the newly established databases.

5. REFERENCES


