International Journal of Computer Assisted Radiology and Surgery (2019) 14:2147–2154
https://doi.org/10.1007/s11548-019-02031-y

ORIGINAL ARTICLE

The development of non-contact user interface of a surgical navigation system based on multi-LSTM and a phantom experiment for zygomatic implant placement

Chunxia Qin[1,2] · Xingchen Ran[3] · Yiqun Wu[4] · Xiaojun Chen[2]

Received: 10 January 2019 / Accepted: 5 July 2019 / Published online: 12 July 2019

Abstract

Purpose Image-guided surgical navigation system (SNS) has proved to be an increasingly important assistance tool for mini-invasive surgery. However, using standard devices such as keyboard and mouse as human–computer interaction (HCI) is a latent vector of infectious medium, causing risks to patients and surgeons. To solve the human–computer interaction problem, we proposed an optimized structure of LSTM based on a depth camera to recognize gestures and applied it to an in-house oral and maxillofacial surgical navigation system (Qin et al. in Int J Comput Assist Radiol Surg 14(2):281–289, 2019).

Methods The proposed optimized structure of LSTM named multi-LSTM allows multiple input layers and takes into account the relationships between inputs. To combine the gesture recognition with the SNS, four left-hand signs waving along four directions were designed to correspond to four operations of the mouse, and the motion of right hand was used to control the movement of the cursor. Finally, a phantom study for zygomatic implant placement was conducted to evaluate the feasibility of multi-LSTM as HCI.

Results 3D hand trajectories of both wrist and elbow from 10 participants were collected to train the recognition network. Then tenfold cross-validation was performed for judging signs, and the mean accuracy was $96\% \pm 3\%$. In the phantom study, four implants were successfully placed, and the average deviations of planned–placed implants were 1.22 mm and 1.70 mm for the entry and end points, respectively, while the angular deviation ranged from 0.4° to 2.9°.

Conclusion The results showed that this non-contact user interface based on multi-LSTM could be used as a promising tool to eliminate the disinfection problem in operation room and alleviate manipulation complexity of surgical navigation system.

Keywords Gesture recognition · Depth camera · Surgical navigation system · Zygomatic implants

Introduction

Image-guided surgical navigation system (SNS) has become an incrementally effected clinical assistance tool for mini-invasive surgery. Since the concept was proposed, this technology has been rapidly developed to apply in various fields, including orthopedics, neurosurgery, otorhinolaryngology and so on. Generally, before the operation, imaging diagnosis with preoperative computed tomography (CT) or magnetic resonance imaging (MRI) was performed to analyze surrounding anatomical tissues and design surgical trajectories by using computer-assisted preoperative planning software. And at the time of surgery, under the guidance of a tracking system, the relative positions among the surgical tools, anatomy structures and planning trajectories could be visualized on a computer screen, guaranteeing the operation accuracy and reliability [2, 3].

✉ Xiaojun Chen
xiaojunchen@sjtu.edu.cn

1 School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

2 Room 805, School of Mechanical Engineering, Shanghai Jiao Tong University, Dongchuan Road 800, Minhang District, Shanghai 200240, China

3 College of Biomedical Engineering and Instrument Science, Zhejiang University, Zhejiang, China

4 Shanghai Ninth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

However, in accordance with the strict operation requirement, all subjects of the surgeon contact must be sterile. However, it is extremely troublesome and time-consuming to disinfect the hardware of the surgical navigation system [4]. Therefore, using standard devices such as keyboard and mouse as HCI is a latent vector of an infectious medium, causing risks to patients and surgeons. Fortunately, three-dimensional hand gesture recognition based on depth camera as an efficient method of touch-free interface has attracted increasing research interests [5, 6]. In general, non-contact hand gesture recognition approaches can be divided into two categories: (1) static hand gesture recognition, which mainly relies on the judgment of difference static hand postures [7, 8]. Unfortunately, this category is infeasible in clinical application as the existence of potential interference of complex surgical postures. (2) Dynamic gesture recognition. Currently, both single poses and continuous multi-label gestures can be distinguished by detecting the begin–end of special gesture from an infinite motion trajectory [9–11]. However, almost all dynamic pose recognition approaches required to abstract the beginning and ending point of special gesture which is a complex task itself.

Therefore, combined with the depth camera, a gesture recognition algorithm was proposed on the basis of an optimized structure of long short-term memory, i.e., multi-LSTM, which allows multiple isolate inputs and takes into account the relationships between inputs layers. The multi-LSTM network was then attached to an in-house oral and maxillofacial surgical navigation system to work as a non-contact user interface. Then a phantom study was involved to evaluate its clinical feasibility and reliability.

Methodology

The architecture of multi-LSTM

LSTM, an optimized network structure of recurrent neural network, can exploit long range relationships in data on the basis of internal purpose-designed memory cells [12, 13]. Figure 1 presents a single-LSTM memory cell, and its data flow can be formulated as:

$$I_t = \delta\left(u_{xi} * x_t + w_{hi} * h_{t-1} + b_i\right) \tag{1}$$

$$F_t = \delta\left(u_{xf} * x_t + w_{hf} * h_{t-1} + b_f\right) \tag{2}$$

$$O_t = \delta\left(u_{xo} * x_t + w_{ho} * h_{t-1} + b_o\right) \tag{3}$$

$$C_t = F_t * C_{t-1} + I_t * \tanh\left(u_{xc} * x_t + w_{hc} * h_{t-1} + b_t\right) \tag{4}$$

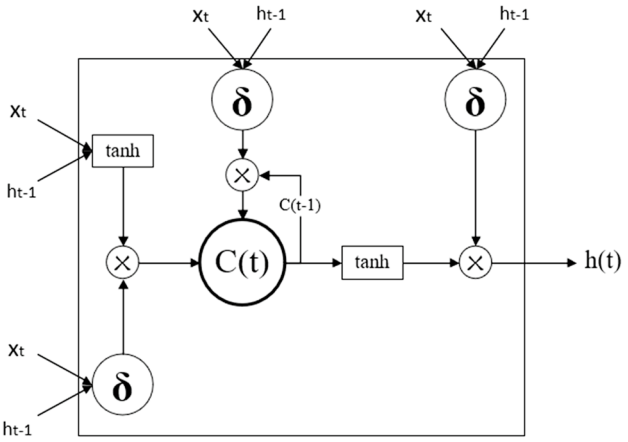

Fig. 1 LSTM cell

$$h_t = O_t * \tanh(C_t) \tag{5}$$

where subscripts t and $t-1$ represent the current and last moment, respectively; δ means sigmoid function; I_t, F_t and O_t are the node value of input gate, forget gate and output gate, respectively; u_{xi}, u_{xf} and u_{xo} are the different weights of input components of different gates; w_{hi}, w_{hf} and w_{ho} correspond with the weights of output in last moment h_{t-1}; C_t represents the status of current memory cell; and h_t means the cell output value.

Figure 2 illustrates the architecture of multi-LSTM, which mainly is comprised of two rows of associated serial LSTM cells. The calculation formulas of $I1_t$, $F1_t$, $O1_t$ and $C1_t$ are same as Eqs. (1)–(4). For the upper row, the calculation function of each node can be presented as follows:

$$h1_{t1} = O1_t * \tanh(C1_t) \tag{6}$$

$$I2_t = \delta\left(q_{hi} * h1_t + u2_{xi} * x2_t + w2_{hi} * h2_{t-1} + b2_i\right) \tag{7}$$

$$F2_t = \delta\left(q_{hf} * h1_t + u2_{xf} * x2_t + w2_{hf} * h2_{t-1} + b2_f\right) \tag{8}$$

$$O2_t = \delta(q_{ho} * h1_t + u2_{xo} * x_t + w2_{ho} * h2_{t-1} + b2_o) \tag{9}$$

$$C2_t = F2_t * C2_{t-1} + I2_t * \tanh\left(q_{hc} * h1_{t1} + u2_{xc} * x_t + w2_{hc} * h2_{t-1} + b_t\right) \tag{10}$$

$$h2_t = O2_t * \tanh(C2_t) \tag{11}$$

$$y = \delta(s1 * h1 + s2 * h2 + b) \tag{12}$$

Compared with Eq. (1)–(5), an additional item $h1_{t1}$ is added in $C2_t$ and the three gates $I2_t$, $F2_t$, $F2_t$. Meanwhile, both outputs of two row $h1$ and $h2$ are contributed to finial prediction.

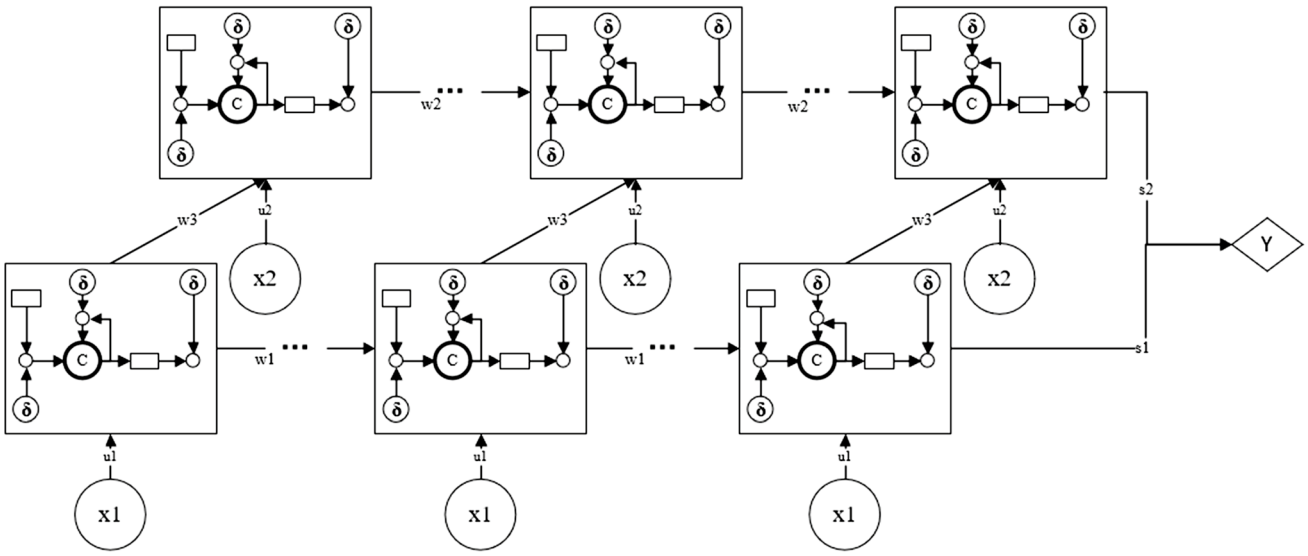

Fig. 2 Architecture of multi-LSTM

Training the multi-LSTM

Data acquisition

In order to combine the gesture recognition with the in-house surgical navigation system of BeiDou-SNS (School of Mechanical Engineering, Shanghai Jiao Tong University) [1], the motion of the right hand is used to control the cursor movement, and gestures of the left hand are designed to manipulate the mouse operation. To investigate the reliability of the gesture recognition algorithm, the trajectory data of both wrist and elbow from 10 participants were collected by a Kinect RGB-depth camera V 1.0 for Windows (Microsoft Inc., USA) to train the multi-LSTM network. During acquisition, the upward direction of the camera was required to align with the operator's vertical orientation, and the imaging plane was adjusted to parallel to the operator's coronal plane. Then, as shown in Fig. 3, the operator cooperated according to the following instructions:

(1) wave upward: firstly, make a wrist–elbow line perpendicular to coronal plane and keep elbow stationary; then, wave hand upward until wrist–elbow line perpendicular to transverse plane; finally, return to the original position;

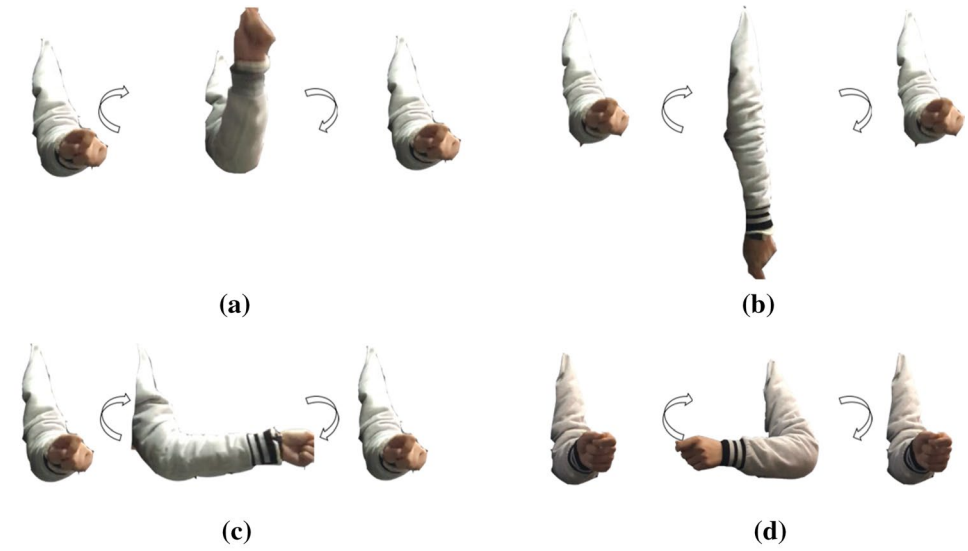

Fig. 3 Gesture schematic. **a** Waving upward, **b** waving downward, **c** waving leftward, **d** waving rightward

(2) wave downward: this process is the same as 1) except for waving downward in the second portion;
(3) wave leftward: this process is similar as 1) except for moving leftward until wrist–elbow line perpendicular with sagittal plane in the second portion;
(4) wave rightward: this process is the same as 3) except for moving rightward in the second portion;
(5) other moving or stationary states: arbitrary movement as long as it is different from the above four categories.

Each participant signed each aforementioned five gestures 50 times, producing 500 specimens for each pose. As little training samples may cause poor performance, massive data are required in our method to train the model appropriately. In order to alleviate this limitation, we applied a data argument method to generate a lot of train data on the basis of our collected data.

Gesture recognition training

As shown in Fig. 4, the red lines are the motion trajectories and the green pots are the coordinate positions of the wrist in different moments. As the velocity variations of hand movement contribute to the maldistribution of the positions of elbow and wrist, a cubic spline interpolation was introduced to preprocess these points before entering the network.

We instantiated the multi-LSTM of network size = 128 to learn the trajectory-to-gesture mapping from interpolated data. The input size was 60 which included 30 wrist input units in x_1 and 30 elbow input units in x_2. As shown in Fig. 4, the output $O = 5$ is the gesture classification which was represented in binary. 00001 to 10000 represent wave upward, downward, leftward and rightward, respectively. All the weights were initialized with sparse connections, and the bias vectors were initialized to 0. In addition, L2 regularization and early stopping were adapted to avoid over fitting. The main parameters are listed in Table 1.

Integration with the surgical navigation system

In order to combine the gesture recognition with the surgical navigation system, aforementioned signs along four directions were designed to correspond to a left button click, right button click, middle wheel forward and middle wheel backward of a mouse event, respectively. Furthermore, the motion of right hand was used to control the movement of the cursor. Theoretically, according to the following equations, the moving vector of the cursor can be obtained by linear mapping from the motion vector.

$$\frac{\partial \vec{d}}{\partial t} = \frac{p_{\mathrm{h}}(t) - p_{\mathrm{h}}(t-1)}{\Delta t} \tag{13}$$

$$P_{\mathrm{c}}(t) = P_{\mathrm{c}}(t-1) + w * \frac{\partial \vec{d}}{\partial t} \tag{14}$$

Table 1 Training parameters used in multi-LSTM network

No.	Parameters	Value	Parameters	Value
1	Learning rate	1.0E−04	Batch size	128
2	Beta 1	0.90	Input size	3
3	Beta 2	0.99	Input units	30*2
4	Epsilon	1.0E−08	Hidden units	128
5	Max iteration	1.0E+07	Classification num	5

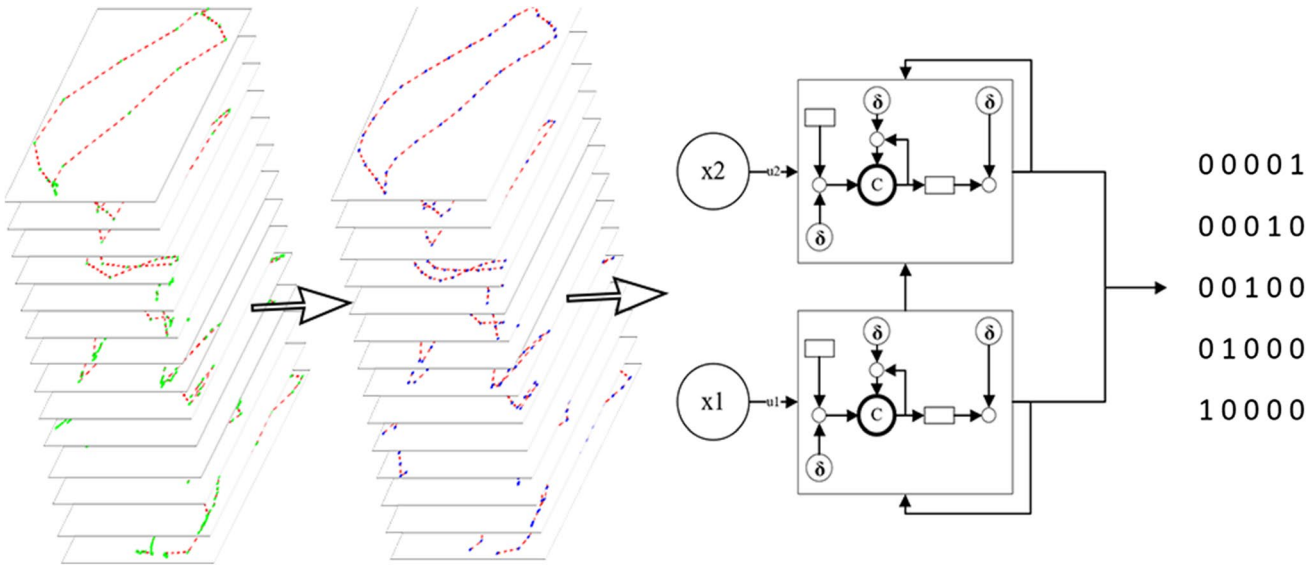

Fig. 4 Process of gesture recognition

where $p_h(t)$ represents hand' position of current moment and $p_h(t-1)$ for the last moment; therefore, $\partial\vec{d}/\partial t$ means the hand' motion vector. w is the mapping factor which is initialized according to the resolution of screen; $p_c(t)$ and $p_c(t-1)$ represent cursor' position of current moment and the last moment, respectively.

However, due to the location draft of depth camera itself and the synergetic effect of limbs, the direct linear mapping results are barely satisfactory. Therefore, a mapping factor which can dynamically be adjusted according to tanh function was adopted.

$$x = \left\| \frac{\partial\vec{d}}{\partial t} \right\| \tag{15}$$

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{16}$$

$$P_c(t) = P_c(t-1) + w * f(x) * \frac{\partial\vec{d}}{\partial t} \tag{17}$$

where x is equivalent to the module of $\partial\vec{d}/\partial t$, that is, the hand' motion velocity.

According to the correspondence between signs and mouse operations, the gesture recognition network was attached to BeiDou-SNS as a sub-thread. And a sliding input model was employed to the recognition approach. As shown in Fig. 5, the latest N sets of data collected from the camera were inputted into the network for each sign judgment. The advantages of this input model are twofold: Firstly, the model effectively eliminates the trouble of query starting point of each gesture. Both isolated gestures and continuous gestures can be recognized. Secondly, the length of input data can be adjustable according to the speed of user movement.

Phantom experiment validation

A phantom study for zygomatic implants (ZIs) placement was conducted to validate the reliability of the HCI of surgical navigation system based on the proposed multi-LSTM. ZI surgery was proposed by Brånemark [14, 15] in 1989 to assist massive grafting surgery or rehabilitate patients who had gone through maxillectomy. As long trajectory is requested in implant embedment, tiny angle deviation or entry point error could lead to intolerable terminal point error [16].

Serving as fiducial makers, eight bone anchored titanium mini-screw (length of 11.0 mm, square cavity of 1.0 mm, diameter of 1.6 mm, CIBEI®, Shanghai, China) were inserted into a resin craniomaxillofacial model. The principle of quantity and distribution of markers was based on the registration criteria [17]. After that, a cone beam computed tomography (CBCT) (Planmeca, Helsinki, Finland) scanning (resolution of 0.33 mm/pixel, slice thickness of 0.4 mm) was performed. Then the CBCT DICOM data were transferred into an in-house oral and maxillofacial planning software [18] and four ZI paths. The resin model and presurgical planning are shown in Fig. 6a, b, respectively.

First of all, the Kinect RGB-depth camera was activated to control the in-house BeiDou-SNS, as shown in Fig. 6f. Then, under the assistance of Kinect RGB-depth camera and the guideline of NDI Polaris (Accuracy of 0.25 mm, Northern Digital Inc., Canada), the phantom experiment of zygomatic implant placement was conducted on a PC with Intel Core I7-7700TM with a 3.60 GHz CPU, 8 GB memory, a 64-bit Windows 10 operating system and a 3 GB NVIDIA GeForce GTX 1060. And the operations were as follows:

(1) Moved right hand and waved left hand upward to open files, including pre-surgical DICOM images, the configuration files of the tracking system and the planning paths;

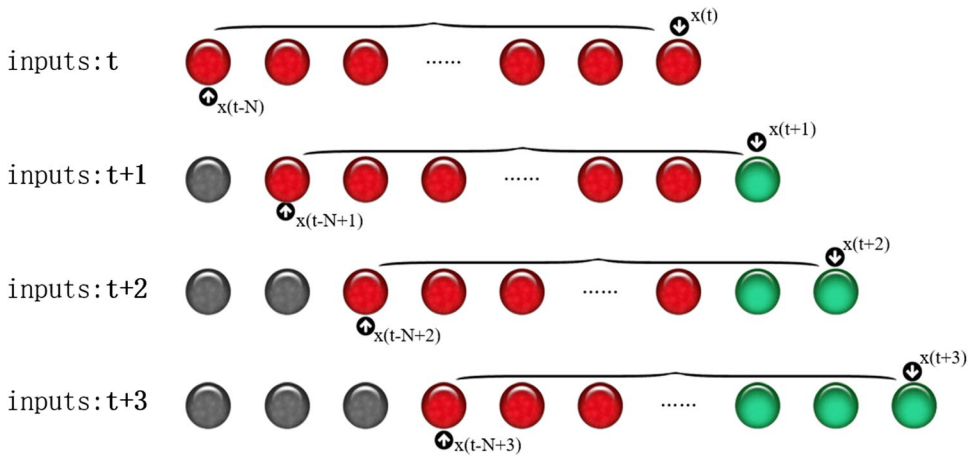

Fig. 5 Sliding inputs of multi-LSTM

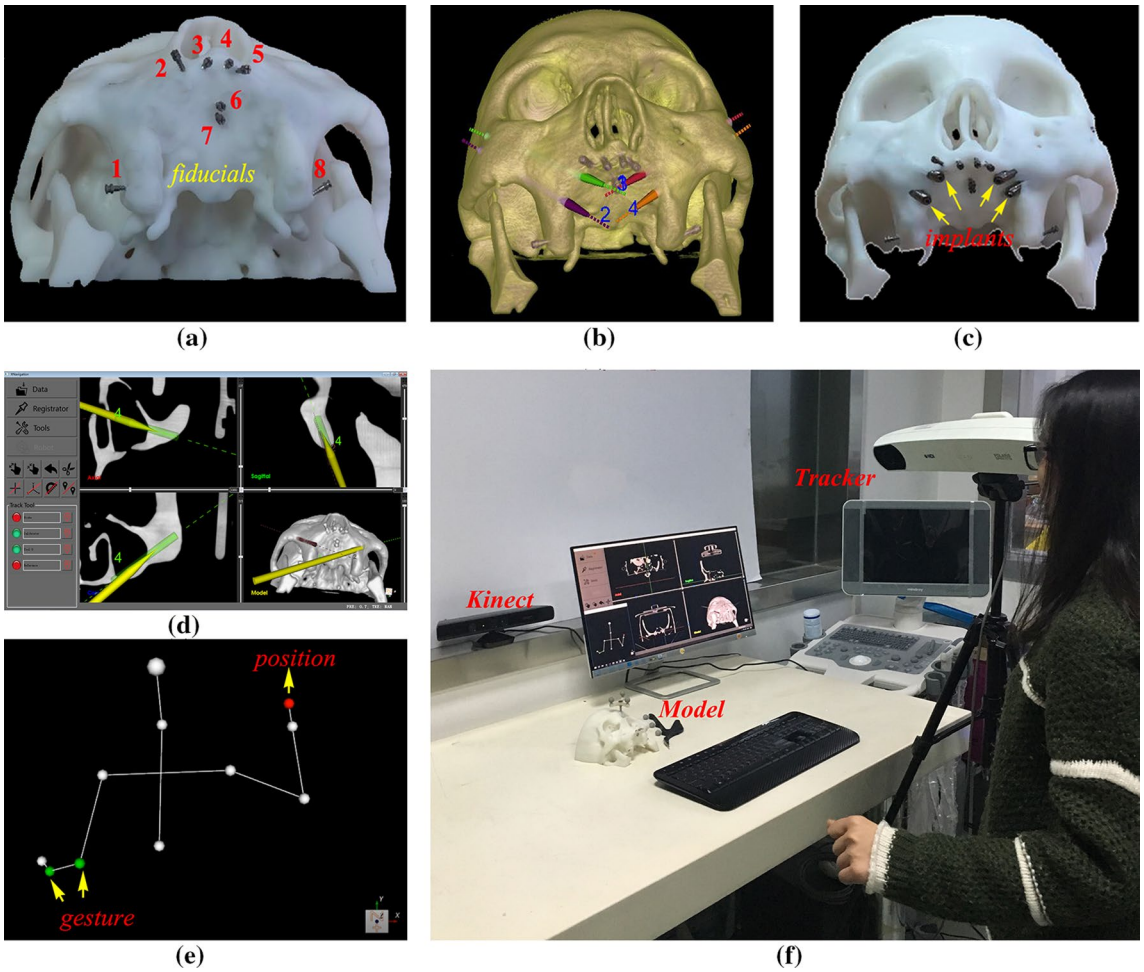

Fig. 6 Procedure and results of the phantom experiment. **a** Preoperative resin model with eight titanium fiducial makers inserted into the maxilla bilaterally; **b** preoperative virtual model and the four zygoma implant paths; **c** postoperative resin model with four implants; **d** intraoperative screen snapshot; **e** real-time skeleton of operator. Different signs were recognized according to the trajectories of left wrist and elbow, and the position of cursor followed the motion of right hand; **f** intraoperative non-contact manipulation of the surgical navigation system via the Kinect RGB-depth camera

(2) Waved left hand leftward or rightward to scan DICOM images by switching the image slice and zooming current image size;
(3) Waved left hand downward to activate the tracking system;
(4) Moved right hand and waved left hand upward to calibrate the surgical instruments and to register the image coordinate space and world coordinate space by starting up corresponding functions.

Results

Recognition accuracy evaluation

As no public dataset meets our train requirement, we recorded 3D hand trajectories of elbow and wrist from 10 participants. Each participant signed each aforementioned five gestures 50 times, providing 500 instances in number for each pose. Three-quarter of the recorded data served as training data and the rest as testing data. Meanwhile, both rotating coordinate and adding noise were used to augment the abundance of train data. To investigate the reliability of the proposed gesture recognition algorithm, tenfold cross-validation was performed at judging signs, and the mean accuracy is $96\% \pm 3\%$.

The results of the phantom experiment

In the phantom experiment, several gestures have been redone because they were not recognized or incorrectly judged sometimes. From statistical data, gestures toward up, down and right three directions could be distinguished with an accuracy of 92%, and the recognition precision of

leftward waves was around 80%. And during the whole experiment, there is no human–computer interaction except the control from Kinect. Along the planned trajectories, four zygomatic implants were successfully placed, as shown in Fig. 6c. After the four implants have been inserted, the 3D model was CBCT scanned again to obtain the postoperative images, which were then fused with the preoperative ones, and three parameters including entry point deviation, exit point deviation and angular deviation were used to evaluate the accuracy of zygomatic implant placement. As shown in Fig. 7, three deviations of four zygomatic implants were measured, and the average deviations of planned–placed implants were 1.22 mm and 1.70 mm for the entry and end points, respectively, while the angular deviation ranged from 0.4° to 2.9°, which can meet clinical requirements. The details are shown in Table 2.

Normally, using a mouse as the human–computer interaction is quite reliable. The default frequency of mouse click for Windows XP is in the range of 1–5 Hz, which depends on the response rate of user-defined double-click events. By comparison, the gesture recognition rate of our system depends on the data update frequency of the RGB-depth camera and the system data acquisition frequency. For different users, the time to complete a wave will be recorded to initialize the user-specific data collection frequency before using the system, ensuring the integrity of the intercepted gesture trajectories. By default, the joint position acquisition frequency of our gesture recognition framework is configured to 60 Hz, and the input of the network requires 30 wrist position units and 30 elbow position units. Therefore, the default recognition rate is 2 Hz, which is slightly slower than using mouse. The maximum recognition rate can be improved by using other RGB-depth cameras with a higher frame rate or using higher frequency of data acquisition.

Table 2 Planned–placed deviation of four implants

No.	Implant length (mm)	Entry point deviation (mm)	End point deviation (mm)	Angular deviation (°)
1	50.00	0.50	1.19	1.39
2	45.00	0.37	1.83	1.88
3	50.00	2.30	1.80	2.9
4	42.50	1.70	1.97	0.4

Discussion and conclusion

As the 3D continuous position of targets can be captured in real time by RGB-D cameras, various gesture recognition approaches have been proposed for HCI, disease detection, robotics and so on [19, 20]. However, as a requisite step in those methods, the algorithm complexity is increased with the distinctions of the beginning and ending points of gestures.

In this study, we proposed an optimized sign judgement structure named multi-LSTM on the basis of traditional LSTM as a method of HCI. To meet clinical requirements, the gesture recognition algorithm was integrated with an

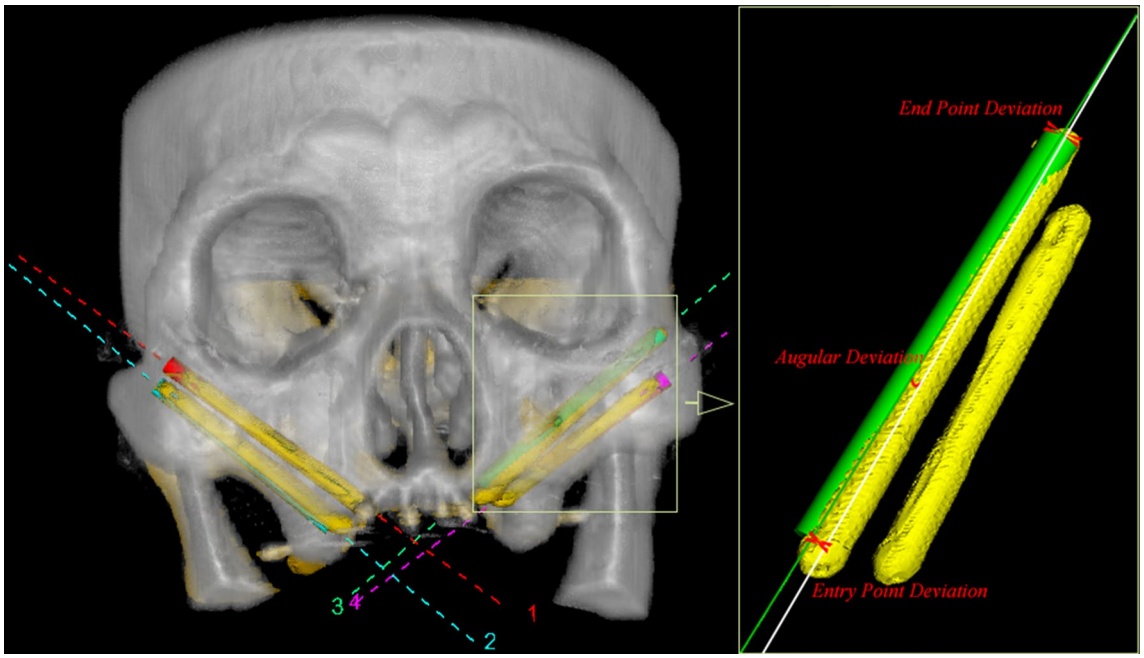

Fig. 7 Fusion of preoperative image and post-operative image and the illustration of planned–placed deviations of implants

in-house surgical navigation system to control the user interface. A phantom study of zygomatic implant placement was conducted to validate its feasibility. As a result, it showed that the non-contact interface based on multi-LSTM could be used as a promising tool to eliminate the disinfection problem for both patients and surgeons.

Although it seems that the results are satisfactory in this study, there are twofold limitations. Firstly, compared to other algorithms, it requires longer time to train the model to suit different users. Nevertheless, it can achieve high speed of gesture recognition in our online test. Secondly, when more than one person is in its detection range, Kinect depth camera will trace several skeletons simultaneously, causing confusion of recognition target. So, user-specific gesture recognition algorithm is expected for further development.

Acknowledgements This work was supported by grants from the National Key R&D Program of China (2017YFB1302903; 2017YFB1104100), the National Natural Science Foundation of China (81828003), the PHC CAI YUANPEI Program (41366SA), the Foundation of Science and Technology Commission of Shanghai Municipality (16441908400; 18511108200), and the Shanghai Jiao Tong University Foundation on Medical and Technological Joint Science Research (YG2016ZD01).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Qin C, Cao Z, Fan S, Wu Y, Sun Y, Politis C, Wang C, Chen X (2019) An oral and maxillofacial navigation system for implant placement with automatic identification of fiducial points. Int J Comput Assist Radiol Surg 14(2):281–289
2. Sukegawa S, Kanno T, Furuki Y (2018) Application of computer-assisted navigation systems in oral and maxillofacial surgery. Jpn Dent Sci Rev 4(3):139–149
3. Chen X, Xu L, Wang H, Wang F, Wang Q, Kikinis R (2017) Development of a surgical navigation system based on 3D Slicer for intraoperative implant placement surgery. Med Eng Phys 41:81–89
4. Ebert LC, Hatch G, Thali MJ (2013) Invisible touch—control of a DICOM viewer with finger gestures using the Kinect depth camera. J Forensic Radiol Imaging 1(1):10–14
5. Cheng H, Yang L, Liu Z (2016) Survey on 3D hand gesture recognition. IEEE Trans Circuits Syst Video Technol 26(9):1659–1673
6. Gkalelis N, Kim H, Hilton A, Nikolaidis N, Pitas I (2009) The i3DPost multi-view and 3D human action/interaction database. In: Proc. conf. vis. media prod, pp 159–168
7. Ren Z, Yuan J, Zhang Z (2011) Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In: Proc. ACM MM, pp 1093–1096
8. Gallo L (2014) Hand shape classification using depth data for unconstrained 3D interaction. J Ambient Intell Smart Environ 6(1):93–105
9. Bhuyan MK, Ajay Kumar D, Macdorman KF, Iwahori Y (2014) A novel set of features for continuous hand gesture recognition. J Multimodal User Interfaces 8(4):333–343
10. Cheng H, Luo J, Chen X (2014) A windowed dynamic time warping approach for 3D continuous hand gesture recognition. In: IEEE international conference on multimedia and expo (ICME)
11. Liou WG, Hsieh CY, Lin WY (2011) Trajectory-based sign language recognition using discriminant analysis in higher-dimensional feature space. In: Proc. IEEE ICME, pp 1–4
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
13. Graves A, Jaitly N, Mohamed AR (2014) Hybrid speech recognition with deep bidirectional LSTM. In: IEEE automatic speech recognition & understanding, pp 273–278
14. Faysal U, Coskun Y, Sener BC, Atilla S (2013) Rehabilitation of posterior maxilla with zygomatic and dental implant after tumor resection: a case report. Case Rep Dent 2013:1–5
15. Aparicio C, Manresa C, Francisco K, Claros P, Alández J, González-Martín O, Albrektsson T (2000) Zygomatic implants: indications, techniques and outcomes, and the zygomatic success code. Periodontology 66:41–58
16. Wang F, Monje A, Lin GH, Wu Y, Monje F, Wang HL, Davó R (2015) Reliability of four zygomatic implant-supported prostheses for the rehabilitation of the atrophic maxilla: a systematic review. Int J Oral Maxillofac Implants 30(2):293–298
17. West JB, Fitzpatrick JM, Toms SA, Maurer CR Jr, Maciunas RJ (2001) Fiducial point placement and the accuracy of point-based, rigid body registration. Neurosurgery 48(4):810–816
18. Chen X, Xu L, Yang Y, Egger J (2016) A semi-automatic computer-aided method for surgical template design. Sci Rep 4(6):20280
19. Bautista MA, Hernandezvela A, Escalera S, Igual L, Pujol O, Moya J, Violant V, Anguera MT (2016) A gesture recognition system for detecting behavioral patterns of ADHD. IEEE Trans Cybern 46(1):136–147
20. Li YT, Wachs JP (2014) HEGM: a hierarchical elastic graph matching for hand gesture recognition. Pattern Recognit 47(1):80–88