Gesture ToolBox: Touchless Human-Machine interface using Deep Learning

Elann Lesnes-Cuisiniez, Jesus Zegarra Flores, Jean-Pierre Radoux

Gesture ToolBox: Touchless Human-Machine interface using Deep Learning

Elann Lesnes-Cuisiniez[1], Jesus Zegarra Flores[1] and Jean-Pierre Radoux[1]

[1] Altran Research, France

Abstract. Human-Computer Interaction (HMI) is useful in sterile environments such as operating rooms (OR) where surgeons need to interact with images from scanners of organs on screens. Contamination issues may happen if the surgeon must touch a keyboard or the mouse. In order to reduce contamination and improve the interactions with the images without asking another team member, the Gesture ToolBox project, based on previous methods of Altran Research, has been proposed. Ten different signs from the LSF (French Sign Language) have been chosen as a way to interact with the images. In order to detect the signs, deep learning methods have been programmed using a pre-trained Convolutional Neural Network (VGG-16). A Kinect is used to detect the positions of the hand and classify gestures. The system allows the user to select, move, zoom in, or zoom out images from organs on the screen according to the recognised sign. Results with 11 subjects are used demonstrate this system in the laboratory. Future work will include tests in real situations in an operating room to obtain feedback from surgeons to improving the system.

Keywords: Human-Computer Interaction · Deep Learning · Kinect.

1 Introduction

Touchless Human-Machine Interface (HMI) is an interdisciplinary field with applications in robotics, computer gaming and sign-language interpretation. Moreover, touchless HMI is very useful in sterile environments such as in the operating rooms (OR) where surgeons need to interact with computers without introducing contamination issues. Most of the time, the joysticks, buttons, or touch screens are wrapped in a plastic and the surgeons need to change their gloves each time they have to use the computers. It is quite common for surgeons to ask colleagues or nurses, who are in another room to interact with the computers for moving images. This does not result in time delays only if colleagues are effectively available [1].

The aim of the Gesture ToolBox project is to propose a simple touchless Human-Machine Interface based on the surgeon's hand gesture recognition using deep learning methods. This investigation is based on previous work by the Altran Research Medic@ team [2] using other machine learning techniques and descriptors of the hand.

2 Related Work

Several projects explore the possibilities of the deep learning method or meeting the needs of the HMI for OR.

Touchless Human-Machine interfaces already exist for surgeons using different techniques. One of these is based on myoelectric signals (MES) [3], unfortunately, it needs electrodes or armbands which are not necessarily comfortable. In another project, L. Di Tomasso, et al. propose a Leap Motion device [4] as a human interface for neurosurgery. There are also solutions on the market for touchless interaction; for instance, the product "Fluid" produced by Therapixel [5]. This solution is based on a depth perception in addition to machine learning techniques that allows pointing one's fingers close to the screen in order to move images. On the other hand, the Gesture ToolBox solution is oriented to interact with cameras standing between one and four meters from the images to be interacted with.

Concerning the deep learning aspect, O. Koller, et al. use a CNN to recognise hand shapes as an example, the main subject of this paper is to combine a CNN and an iterative EM algorithm to train the CNN on a big dataset weakly labelled [6]. Another paper from Huang, et al. describes the research of finger key point's detection from a mobile camera [7]. Their system is robust to changing background, however it is available for only one finger which is not sufficient in the context of the Gesture ToolBox project. Two other projects use deep learning algorithms and the Kinect for hand segmentation and tracking [8] or sign language recognition [8, 9]. L. Pigou, et al. include the recognition of the body and descriptors [8]. In the method, presented in this paper, there is no hand feature extraction; this was a major part of the machine learning based method mentioned in [2].

3 Methodology

3.1 The Ten Hand Gestures

In order to test the hand gesture recognition, ten gestures were chosen (Fig. 1) from the French sign language (LSF). The algorithm has been trained with these particular gestures because of their simplicity (not causing additional fatigue to the surgeons).

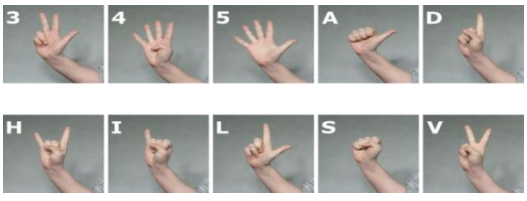

Fig. 1. Hand postures of the ten gestures

3.2 Gesture ToolBox System

Users stand in front of the Kinect and perform the defined gestures with their right hand in order to manipulate images on the screen in real time. The five main steps of the program (Fig. 2., on the left) are repeated in every frame. The upstream training phase of the neural network runs once. The code is flexible enough to recognise different signs without major modifications thanks to the ease of use of this deep learning approach.

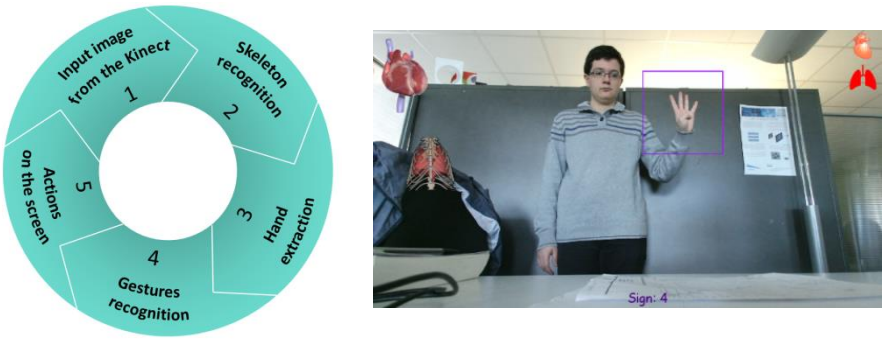

Fig. 2. Architecture of the gesture recognition system (left) and user interface (right)

Input Image from the Kinect and Skeleton Recognition. It uses the Microsoft Kinect for Windows V2. This device is able to track people and their skeleton (up to 25 skeletal joints of a maximum of 6 people) [10].

Hand Extraction and Gesture Recognition. Once the skeleton is identified by the Kinect (Fig. 2.), the position of the right hand is extracted for every frame; a picture centred on the entire right hand is obtained.

Before launching the program, two files which contain the structure and the weights of the neural network trained to classify the ten gestures, are loaded. The neural network is fed every extracted RGB picture and classifies the gesture. In order to reduce imprecisions, the result displayed is the most represented sign among the last five classifications.

Actions and User Interface. The detected gestures are used to select, move, zoom in or zoom out the images of the heart and the lungs on the screen. Future applications will include moving real medical images or specific 3D objects from the industry.

3.3 Deep Learning

The gesture recognition phase of the project is done by a convolutional neural network (CNN) adapted to the classification of pictures.

Data Acquisition. The project functions in real time dealing with pictures from videos. Data acquisition must be specific for each type of gesture in each type of environment (laboratory, very bright OR, etc.). For proof of concept, pictures were collected from eleven people with different skin colours in front of a metallic closet which provides a bicolour background. To simulate new pictures in order to increase the size of the dataset, a small random translation and rotation was applied.

Classification into ten classes is a supervised problem, consequently, a label was placed on the corresponding pictures. The final training dataset contains more than 2600 pictures for the ten classes with a quantity of between 220 and 320 pictures for each class. This remains a small dataset; as a consequence, much attention was given to the issue of overfitting.

Transfer Learning. A pre-trained neural network was used, in our case, the VGG-16 neural network [11] already trained on the ImageNet dataset [12] has been chosen and retrained with our training dataset.

4 Tests and Results

The tests are conducted having the same background as in the training phase. In the early state of the investigation, only four people contributed to these tests, including one subject who did not contribute to the data acquisition.

4.1 Confusion Matrix

The confusion matrix (Fig. 3.) provides the results of the 333 gestures done.

	3	4	5	A	D	H	I	L	S	V
3	20	0	0	0	0	0	0	1	1	0
4	0	34	1	0	0	0	1	1	0	4
5	3	0	39	1	0	0	0	0	0	0
A	0	0	0	39	0	0	0	2	2	0
D	0	0	0	0	26	0	0	1	0	3
H	0	0	0	0	4	26	2	0	1	3
I	0	0	0	0	2	2	18	2	1	0
L	4	0	0	1	0	0	0	26	0	0
S	0	0	0	2	4	0	0	1	29	0
V	2	0	0	0	2	0	0	0	1	21

Fig. 3. Confusion matrix (left) and relative frequencies of the classes (right)

4.2 Evaluation

Thanks to the confusion matrix (M_{ij}), the precision, recall, and the f-score for each sign were obtained (Table 1.).

Table 1. Precision, recall, and f-score

	Precision $(\frac{M_{ii}}{\sum_j M_{ji}})$	**Recall** $(\frac{M_{ii}}{\sum_j M_{ij}})$	**f-score** $(2 \times \frac{precision_i \times recall_i}{precision_i + recall_i})$
3	69.0 %	90.9 %	**78.4 %**
4	100.0 %	82.9 %	**90.7 %**
5	97.5 %	90.7 %	**94.0 %**
A	90.7 %	90.7 %	**90.7 %**
D	68.4 %	86.7 %	**76.5 %**
H	92.9 %	72.2 %	**81.3 %**
I	85.7 %	72.0 %	**78.3 %**
L	76.5 %	83.9 %	**80.0 %**
S	82.9 %	80.6 %	**81.7 %**
V	67.7 %	80.8 %	**73.3 %**
mean	**83.1 %**	**83.1 %**	**83.1 %**

5 Discussion

It is important to say that the program was deliberately given in difficult situations in order to test its limits: people very far from the Kinect, very close, far from centre, under a very strong light, with the right hand in front of the head or body. If the user is at a correct distance and without excessive light or lack of light, the program has fewer errors. Consequently, the authors would like to point out two main aspects.

5.1 Kinect's Limitations

These results do not take into account bad skeleton recognition from the Kinect. Sometimes, the Kinect is not able to detect the skeleton or distorts it. As a consequence, it does not place the right hand at the correct position. In such cases, the last known position of the right hand is used in order to extract the current hand position. In most of the cases, it is a good approximation because the user does not move his or her hand very abruptly. In other cases, the only solution is asking the person to move.

5.2 Errors

Three most common errors have been observed:

- The neural network is confused by two very similar signs. For instance, it confuses the “H” and “V” signs if the users have a small gap between their index finger and their middle finger. (Fig. 6).
- The neural network does not “see” one or two finger(s). Sometimes, the neural network transforms “4” into “V” or “5” into “3. In the similar way, it transforms “H” into “D” (and sometimes, “H” into “I”).
- The neural network “sees” one additional finger. Sometimes, it has been observed that it classifies a “D” into a “V” or the sign “L” into a “3” because it “adds” a finger near the others.

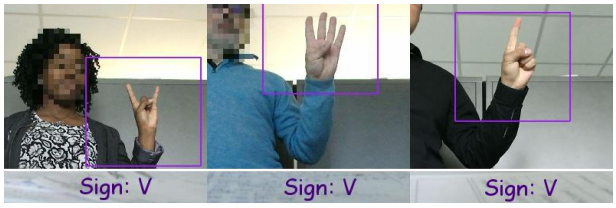

Fig. 6. Intended sign: “H”, recognised sign: “V” (left), intended sign: “4”, recognised sign: “V” (middle), intended sign: “D”, recognised sign: “V” (right)

6 Conclusions and Future Work

In this paper, a deep learning solution for HMI was presented. The goal was not to prove that deep learning method obtains better results than other solutions, in particular classical machine learning methods, but to propose another way to process gestures. In a previous work done by Altran research, Belhaoua, et al. [2], hand-crafted features were computed and decision trees were used for the classification. However, it might be interesting to mix the deep learning approaches and more classical methods of image processing or machine learning in order to overcome the Kinect’s limitations and resolve the most commonly observed errors.

Tests in real conditions in operating rooms (OR) are now necessary in order to take into account the surgeons’ feedback to improve the user interface to fill their requirements. The creation of an interface which allows them to register their own gestures in their particular environment and use them in the touchless interface is already implemented.

More data will be necessary in order to reduce the defect of overfitting. Future improvement may include the addition of the depth and infrared values provide by the Kinect to a neural network using transfer learning. To give more functionality to the surgeon, we may explore detecting both hands of the user using the mirror image of the right hand. Finally, we may consider the use of standard cameras.

References

1. O'Hara, K., Sellen, A., Gonzalez G., Carrell T., Mentis H., Penney G., Criminisi A., Corish R., Rouncefield M., Dastur N., Varnavas A.: Touchless interaction in surgery. In: Communications of the ACM, 57(1), 70-77 (2014).
2. Belhaoua A., Krebs A., Radoux, J-P.: Gesture-Based Interaction on Surgical Field Using Touchless Technology. In: Conférence Reconnaissance de Formes et Intelligence Artificielle on 2016,
3. Hettig J., Mewes A., Riabikin O., Skalej M., Preim B., Hansen C.: Exploration of 3D Medical Image Data for Interventional Radiology using Myoelectric Gesture Control. In: 7th Eurographics Workshop on Visual Computing for Biology and Medicine (2015)
4. Di Tommaso L., Aubry S., Godard J., Katranji H., Pauchot J.: A new human machine interface in neurosurgery: The Leap Motion®. Technical note regarding a new touchless interface. Neuro-Chirurgie (62:3), 178-181 (2016)
5. Therapixel Homepage, http://www.therapixel.com/, last accessed 2017/05/03
6. Koller O., Ney H., Bowden R.: Deep Hand : How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3793-3802, Las Vegas, NV, USA, June 2016.
7. Huang Y., Liu X., Jin L., Zhang X.: DeepFinger : A Cascade Convolutional Neuron Network Approach to Finger Key Point Detection in Egocentric Vision with Mobile Camera. In: Systems, Man and Cybernectics (SMC), IEEE International Conference on (2015)
8. Pigou L., Dieleman S. Kindermans P-J., Schrauwen B.: Sign Language Recognition using Convolutional Neural Networks. In: Agapito L., Bronstein M., Rother C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science, vol 8925. Springer, Cham
9. Tang A., Lu K., Wang Y., Huang J., Li H.: A Real-time Hand Posture Recognition System Using Deep Neural Networks. ACM Trans. Intell. Syst. Technol. 9, 4, Article 39 (2013)
10. Developing with Kinect for Windows Homepage, https://developer.microsoft.com/en-us/windows/kinect/develop, last accessed 2017/07/03
11. Simonyan K., Zisserman A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556v3
12. ImageNet Homepage, http://www.image-net.org/, last accessed 2017/07/03