Gesture Recognition with Microsoft Kinect Tools for Socially Assistive Robotics Scenarios

Paulo Henriques
paulo.henriques@tecnico.ulisboa.pt

Supervisors: Alexandre Bernardino & José Santos Victor
Instituto Superior Técnico, Lisboa, Portugal

***Abstract*— Vizzy is a socially assistive robot being developed in Instituto Superior Tcnico to work as a personal trainer for elderly, disabled and injured people in clinics, nursings, etc. A Human-Robot Interaction (HRI) System based on gesture recognition is proposed. The system is being developed to allow people with some physical limitation give instructions to the robot. A Kinect II was used together with Visual Gesture Builder to allow gesture recognition. Two detections filters were implemented to upgrade the behavior of the classifiers and allow the users to have a more enjoyable experience.**

Keywords: Augmented Human Assistance, Social Robot, Gesture Recognition, Human-Robot Interaction

I. INTRODUCTION

Augmented Human Assistance(AHA) is a project from a multi-disciplinary consortium and intends to fight problems associated with aging and sedentary lifestyle. Vizzy is a being developed under the AHA project to work as a personal trainer for people with personal necessities (elderly, injured, etc.).

The present work aims to develop a human-robot interaction system based on gestures performed by the user. This group of people can feel some difficulties dealing with technology, this challenged us to design and develop a more natural and easy to use interface for this population.

The recognition system intends to deal with two kinds of gestures: command and status gestures. The first ones were designed to give specific instructions to the robot, for example, to call it, and the second ones to give feedback to the robot, for instance, ask it to repeat something.

To Vizzys success, it has to have a pleasurable way to interact with people, and this work aims to contribute to it. We developed an entirely functional communication system that allows people to give instructions to Vizzy using natural gesture language.

II. REALATED WORK

The interaction between human and robots has been evolving since the early days, from a command line to solutions with voice or gesture recognition. Gesture recognition can be performed with solutions based on thresholds that can fail quickly, because of the variety of dimensions that users can have; or with solutions based on machine learning where a wide range of samples can be used and improve the adaptability of the system.

In [1] is considered a scenario where someone has to drive a car using gestures, to study the intuitiveness of gestures used by individuals to complete determined actions. Authors developed a method to quantify the intuitiveness of each gesture associated with navigation actions. In the end, they observed that in a universe of 38 men they created the 280 different actions for the same 8 actions.

In [2] it is exploited the application of gesture-based interaction in sites like museums. The authors developed projects that were implemented in Vatican Museum and allowed to understand that several factors can influence the intuitiveness and naturality of gestures such as nationality, short usage time, limited space. This project allowed to discover that people from with the same nationality do not associate the same gesture to an action, but patterns can be found.

In [3] authors test three different ways to guide a robot to understand which one requires less workload to the user. The three methods of analysis were: 1) Direct Physical Interaction (DPI), where people use hands to guide the robot but the motors do the force required to the movement; 2) Person Following, here the robot follows the user; 3) Pointing Control, users use their arms to point the site where the robot should go. In the end, authors believe that DPI is the best interaction modality, but when hands of the operator are busy, the person following algorithm is the ideal method.

The work [4] proposes and studies the use of two sensors in the interaction between humans and a humanoid robot. The sensors used were a Kinect I and a CyberGlove II, and allowed researchers acquire a bundle of 12 human upper body gestures. The authors constructed a gesture database with 25 people with different body size, gender, and cultural background. In the end, they obtained more than 96% of accuracy when the system was trained with a sufficient number of samples.

In [5] a Human-Robot Interaction (HRI) system was developed to integrate into flying robots. One of the principal problems in gesture-based HRI systems is the individuality of subjects. Each person has his natural way to execute

predefined gestures. Authors of the paper studied the possibility to use a transfer learning algorithm, training a classifier with gestures from a database and gestures provided by the user of the robot. The results of the experiments allowed to understand that when the classifier is trained using simultaneously concrete samples from the user and samples from other sources, it can be more efficient turning into a more reliable machine.

In [6] researchers focused on the detection of gestures that have two characteristics: regularity and repeatability. Researchers used Kinect I to track human skeleton with a speed of 30 frames per second(fps). From data recorded by the camera, selecting specific joints and measuring movement angle during a period, a sinusoidal signal is obtained. With knowledge of which joints are producing the wave signal, and with mathematical information of the signal, gestures can be recognized. The recognition system was designed using a multi-layer perceptron (MLP). The MLP was trained using a bundle of 8 different gestures: Hi, Come, Clapping, Driving, Hungry, O, Stop, Noisy. The entire System can be described in three steps: 1) pre-processing, the graphic result is converted to numerical data; 2) model comprehension, data is categorized and comprehended in the MLP; 3) Gesture Recognition. Good results were obtained with this methodology, 2 gestures of 8 were recognized by the system all time they were performed. The minimum accuracy value was 73.3%, associated with the gesture "Come". In general, the average accuracy of the system was 85.8%.

In [7], researchers developed a robot to help people working out and to engage elderly people in physical activity. This robot was designed to promote upper body exercise and has capabilities to demonstrate exercises with mechanical arms.

Researchers performed tests to try to discover the benefits of using a relational robot vs. a non-relational robot. Users worked out with both types of robots, and the results showed that relational robot received high scores by the majority, according to levels of enjoyableness and usefulness.

To evaluate the interaction preference between a physical robot and a simulated robot, seniors worked out with a real robot and with a simulation of a real robot. In the end, users evaluated the sessions with the physical version as more enjoyable and useful. Users referred that the physical robot was more socially attractive and it provided a better sense if social presence.

III. Methods

Visual gesture builder(VGB) is a software that integrates the Kinects SDK and was designed to generate gestures databases. The gesture database can be used by applications to perform real-time gesture detection [8]. To generate these databases it has to perform machine learning processes, VGB includes two different software able to perform machine learning processes, AdaBoostTrigger and RFRProgress. For this work, we used the first option.

The gesture database identifies gestures when a user performs a gesture, the database generates signals with intervals of milliseconds, so when a gesture is performed during some seconds, the database produces tens of gesture events. To solve this overflow two versions filters were developed, first one giving priority to the accuracy and second one giving priority to the speed.

A. Data Aquisition

Our datasets for training and testing were recorded with a Kinect II and using the software Kinect Studio V2.0. While recording video, the system was configured to record the skeleton and depth information, at 30 frames per second(FPS).

VGB provides an interface that allows for training and testing classifiers. To train the classifiers we had to tag, in each video, what was part of the gesture and what was not part of the gesture. Each gesture has a correspondent classifiers. Each classifier was trained with positive samples of the associated gesture and at same with negative samples of all other gestures.

Adaboost is the basis algorithm to the machine learning part of this work. In this algorithm, the output of other learning algorithms, weak learners, is combined aggregated in a weighted sum that represents the final output of the classifier.

B. Fast Filter

The second version of the filter was designed just to avoid overflow of information, contrarily to the first version. For this version just one database was required: Vizzy - where gestures are saved, gestures present in the database are responsible for the actions of the robot. After a gesture is sent to a decision module, the system has a pause of 1.8 seconds.

C. Slow Filter

The first version was developed to avoid overflow and to improve the accuracy of the system. This version requires that the user performs the gesture during 2 seconds and the system processes all gestures detected during those 2 seconds to decide which gesture may be recognized by the robot. For this version a database and a buffer were used: 1) Buffer to save all gestures identified the last two seconds, with a capacity for 8 gestures; 2) Database where gestures were collected after being processed, and are responsible for the actions of the robot. Gestures are saved first in the buffer with a maximum frequency of 4 gestures per second. Every time that the database doesnt detect activity for more than 2 seconds the buffer is dropped. Whenever the time difference between the first and the last entry in the buffer

is higher or equal to 2 seconds, a gesture is sent to decision algorithm, which scans all gestures in the buffer and sums the confidence of repeated gestures. In the end, it picks the gesture with more confidence and publishes it in the Vizzy database.

IV. Experimental Setup

The current work is being developed to integrate into a humanoid robot.It uses a set of sensors to allow the correct operation of the robot.

A. Kinect II

Kinect II can provide depth and RGB data with a frame rate of 30fps, and tracks up to 6 people simultaneously with 25 joints per person. Furthermore, it is also equipped with Infrared(IR) capabilities and 4 microphones to capture sound. Kinect gives the guarantee of accurate measures in a range of 0.5 meters to 4.5 meters.

B. Datasets

The machine learning algorithm used requires the usage of datasets, for training and testing, on video format. The training set was recorded with a group of 7 people with different body size, gender, cultural background, aged between 23 and 50. The test set was recorded with a different group, composed of 2 people with different body sizes, males and aged between 24 and 25.

C. Gestures

The gestures were designed to allow users to give instructions and feedback to the robot. They can be grouped into two different groups: Static and dynamic gestures. Static gestures are: ”Don’t understand”, ”have a question”, ”Ok 1”, ”Ok 2”, ”phone call”, ”confident 1”, ”confident 2”, ”come here”, ”go out” and ”stop”. Dynamic Gestures: ”Next exercise”, ”rotate”, ”finish session”, ”calm down”, ”pay attention” and ”bye bye”.

TABLE I: Static Gestures

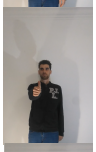
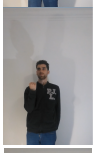
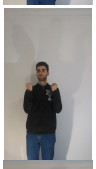
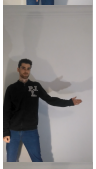
No.	Gesture	Meaning
1		Dont understand
2		Have a question
3		Ok 1
4		Ok 2
5		Phone call
6		Confident 1 hand
7		Confident 2 hands
8		Come here
9		Go out
10		Stop

TABLE II: Dynamic Gestures

No.	Gesture	Meaning
11		Next exercise
12		Rotate
13		Finish session
14		Calm Down
15		Pay Attention
16		Bye Bye

D. Software platform

Besides the software available from Kinect we developed an interface to connect the user and the robot. The interface was designed with C#, and it has capabilities to track up to 6 users simultaneously. This interface allows the users to have instantaneously feedback if they are being monitored if the gesture is being detected and which gesture is being detected. In the back-end, we created a MongoDB database to save gestures detected, which is read by the state machine of the robot and allow it to act accordingly with the input provided by users.

E. Validation and tests

To test the entire system, including filters, a series of tests with users were done. These tests were performed with a group of eight persons, from different backgrounds, different genders, body dimensions, and between 23 and 52.

A software tool was developed to evaluate the satisfaction levels of the users when dealing with the gesture recognition tool and to compare both filters described in section III. This tool associated each gesture to a letter or character, as shown in Table III, and the users were asked to write a complete sentence, I am here trying the new system!. While they were performing the experience, they had visual access to what they were writing on a screen, and the visual information was equivalent to what is presented in figure 1. Each user was asked to write the sentence two times, each one with a different filter. Half of the users tried the fast version first, and the other half tried the slow version first, this strategy was implemented to avoid a potential habituation effect experienced by the users that could affect the final results. In the end, users were invited to fill a questionnaire to evaluate the experience with each one of the versions.

TABLE III: Affirmative sentences present in the survey.

Gesture	Character
Calm Down	A
Come Here	E
Confident 1	!
Confident 2	G
Don't Understand	H
Finish Session	I
Go Out	M
Have a Question	N
Next Exercise	R
Ok 1	T
Ok 2	W
Pay Attention	Y
Phone Call	"Space"
Rotate	S

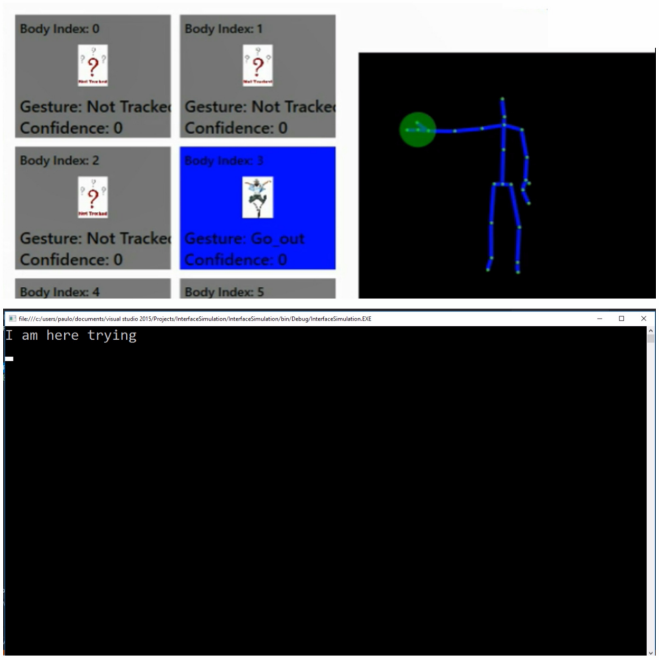


Fig. 1: On top image is shown the interface where the user can see which is the gesture being identified and above the console where the sentence was being displayed.

F. Robot Integration

In the end, the developed software was integrated into the robot. At the time of the integration, the internal robot

software had a state machine ready to receive gesture events and perform accordingly with gesture received. At that moment robot was limited to 4 head movements, so we associated each gesture to a head movement. The camera used was not integrated into the robot, so we established wi-fi communication between Vizzys computer and computer where the camera was connected. The detection software was performing real-time detection and publishing gestures in a local MongoDB database. Vizzys computer was reading the database and sending the signals to the state machine. A small sequence of gestures and response of the robot can be observed in figure 2.

Fig. 2: Robot integration, sequence of gestures being executed and head response of the robot

V. Results

We have performed two types of tests: 1) Evaluation of classifiers, to evaluate the performance of the classifiers after training; 2) Evaluation with users to evaluate the degree of satisfaction of users when dealing with both versions of detection filters.

A. Classifiers Evaluation

Confusion Matrix: A confusion matrix, presented in table IV, is used to gather information about the classification performance of the developed classifiers. It is a two-dimensional representation, where one dimension represents the true class of a sample, and the other represents the classification assigned by the classifier. In a confusion matrix is possible to identify four different categories: True Positives(TP) that correspond to samples correctly classified as positives; False Positives (FP) corresponding to negative samples classified as positive; True Negatives (TN) corresponding to negative examples classified as negatives; False Negatives (FN) that correspond to positive samples classified as negative. The test set was recorded with 2 different users. Each one was asked to perform each gesture individually 2 times. In the end, the test set was composed of 4 samples of each one of the gestures. All this means that ideally, the confusion matrix would be a diagonal matrix with 4 in the main diagonal.

TABLE IV: Confusion Matrix

Gesture \ Classifier	Bye bye	Calm down	Come here	Confident 1	Confident 2	Don't understand	Finish session	Go out	Have a question	Next exercise	Ok 1	Ok 2	Pay attention	Phone Call	Rotate	Stop	Non-Detected
Bye bye	**0**	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	4
Calm down	0	**3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Come here	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Confident 1	0	0	0	**3**	0	0	0	0	0	0	0	0	0	0	0	0	1
Confident 2	0	1	0	4	**4**	0	0	0	0	0	0	0	0	0	0	0	0
Don't understand	0	0	0	0	0	**2**	2	0	0	0	0	0	0	0	0	0	2
Finish session	0	3	0	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0
Go out	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	1
Have a question	0	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	1
Next exercise	0	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0
Ok 1	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	3
Ok 2	0	0	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	1
Pay attention	0	0	0	0	0	0	0	0	0	0	0	0	**3**	1	0	0	1
Phone Call	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0	0
Rotate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0
Stop	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	**3**	1

Precision and Recall: Precision, shown in Table V, is a measure of the result relevance and recall, shown in Table VI, relates to the number of relevant results returned. A classification system with high recall and low precision has a high number of detections, but most of them are misclassified. A classification system with great precision and reduced recall detect a small number of samples, but most of them are correctly identified. Ideally, a system is expected to have high recall and precision, that would mean that system is performing accurate detection and at the same time returning the majority of all positive results.

F1-Score: In the field of the binary classification F1 score (see Table VII) also known as F-score or F-measure, works as a measure of the accuracy the system under test. It considers the precision and recalls to the computation of the scores. It can be interpreted as a weighted average of recall and precision. In the ideal case, F-scores tend to 1, and the worst case occurs when they tend to 0. The F1 score is computed as a harmonic mean of precision and recall multiplied by 2, to make the value equal to 1 when recall and precision are both 1.

TABLE V: Precision

Classifier	Bye bye	Calm down	Come here	Confident 1	Confident 2	Don't understand	Finish session	Go out	Have a question	Next exercise	Ok 1	Ok 2	Pay attention	Phone Call	Rotate	Stop
Precision	0	0.428571	1	0.428571	1	1	0.666667	1	1	0.444444	1	1	1	0.8	1	1

B. Choice of the best classifiers

"Bye Bye" is clearly recognized as the worst gesture of the batch, since the classifier had no capabilities to recognize any

TABLE VI: Recall

Classifier	Bye bye	Calm down	Come here	Confident 1	Confident 2	Don't understand	Finish session	Go out	Have a question	Next exercise	Ok 1	Ok 2	Pay attention	Phone Call	Rotate	Stop
Recall	0	0.75	0.5	0.75	1	0.5	1	0.75	0.75	1	0.25	0.75	0.75	0.25	1	0.75

TABLE VII: F-Score

Classifier	Bye bye	Calm down	Come here	Confident 1	Confident 2	Don't understand	Finish session	Go out	Have a question	Next exercise	Ok 1	Ok 2	Pay attention	Phone Call	Rotate	Stop
F-Score	0	0.545	0.667	0.545	1	0.667	0.800	0.857	0.857	0.615	0.400	0.857	0.857	0.381	1	0.857

samples during the training stage. This gesture is periodic and the number of periods is not defined, which means each user performs it in a different way. This fact combined with the small number of samples can cause the lack of performance of the classifier. This gesture was not included in the test stage with users.

Considering F-score, because it takes into account recall and precision, it is perceptible that "Ok 1" and "Phone Call" have both low performance. The first one is what is considered a small gestures, with small amplitude of movement and with only a small part of body involved. In general, this type of gestures needs a bigger number of samples in training set to present similar results when compared with big gestures.

The gestures with the best f-scores are "Confident 2", "Ok 2" and "Pay attention". The former achieves a perfect score. Since it is a simple and big gesture, both facts combined can contribute to a low number of samples required in the training stage to achieve a good result. This result is inflated by the reduced number of samples, used in training and testing stages. With a higher number of samples, the result would not be exactly 1 but probably around 1. The good result also achieved by the gesture "Pay attention" can be explained by the same reasons that were pointed for the previous gesture. In this batch "Ok 2" can be considered an outsider because its characteristic and similarity with gestures like "Ok 1" and "Stop" would suggest that a huge number of samples would be required to achieve a good result.

C. Results of evaluation with users

Vizzy is being developed to work directly with people so tests with users can have a huge relevance to the global evaluation of the system. As explained in previous sections, a group of users was asked to test the system and fill a survey with questions about the interaction. The survey was composed of 8 affirmative sentences, for each one of the filter version, and users were asked to evaluate each sentence on a scale of 1 to 5. The affirmative sentences are present in Table VIII.

TABLE VIII: Affirmative sentences present in the survey.

Nr.	Question
I	I was able to write the entire sentence.
II	It was easy to write the sentence.
III	It was easy to correct the errors.
IV	It was easy to deal with the delay.
V	It had an adequate response time.
VI	The system identified each gestures instantaneously.
VII	It had a low number of misidentified gestures.
VIII	It was an engaging system.

Results show that users prefer the interaction with the fast version of the filter. The referred version received always higher evaluations when the question asked about positive characteristics, and lower evaluations when questions asked about negative aspects. Considering only the last question, the one that evaluates the overall system, we can observe a gap between the users opinion: The fast but less precise version receives 4.83 and the slow but more precise version receives 3.33. The preference presented by the users can be influenced by the fact that the version 2 doesnt need that users hold the gesture so much time when compared with the other version. This fact makes the algorithm detect more inputs, correctly or wrongly, but allowing at the same time easier and faster corrections. During the test phase, it was possible to observe that version 1 due to the delay in the detection of visual inputs gives the feeling that the gesture is not being detected, this makes the users give up before the detection of the gesture. Due to this fact was possible to detect signals of frustration when dealing with the software.

VI. Conclusion

In the present work, we were able to implement and also integrate on Vizzy capabilities to recognize gestures performed by users and to act accordingly with the gestures.

The recognition system has shown good general performance dealing with this kind of situations. However, it has shown difficulties dealing with periodic gestures, where the number of periods is not well defined. It also showed a better performance dealing with gestures involving bigger parts of the human body.

After submitting both filters to functional tests was possible to conclude that the fast version had a better acceptance by the users. It has less precision but the fast response allows users to have a more pleasurable experience.

A. Future Work

Despite we have achieved reasonable results is important to not forget the objective of the work developed: to be applied in a real robot, that is being designed to work with real users. So, the performance must be as good as possible. Improvements can be achieved if the number of samples used during the training stage increases, the performance can also

increase if the system is trained with samples recorded with people belonging to Vizzy's target audience. Would also be positive to ask the same group of people to evaluate the satisfaction using the system.

REFERENCES

[1] H. I. Stern, J. P.Wachs, and Y. Edan, Optimal consensus intuitive hand gesture vocabulary design, The IEEE International Conference on Semantic Computing, pp. 98103, 2008.

[2] S. Pescarion, E. Pietroni, L. Rescic, M. Wallergard, K. Omar, and C. Rufa, Nich: a preliminary theoretical study applied to cultural heritage contexts, IEEE - Digital Heritage International Congress (DigitalHeritage), pp. 355 362, 2013.

[3] A. Jevtic, Y. Parmet, and Y. Edan, Comparison of interaction modalities for mobile indoor robot guidance: Direct physical interaction, person following, and pointing control, Human-Machine Systems, IEEE Transactions on (Volume:45 , Issue: 6), pp. 653 663, 2015.

[4] Y. Xiao, Z. Zhang, A. Beck, J. Yuan, and D. Thalmann, Human-robot interaction by understanding upper body gestures, PRESENCE TELEOPERATORS & VIRTUAL ENVIRONMENTS, vol. 14, no. 3, pp. 133154, 2014.

[5] G. Costante, E. Bellocchio, P. Valigi, and E. Ricci, Personalizing vision-based gestural interfaces for hri with uavs:a transfer learning approach, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 33203326, 2014.

[6] J.-H. Choi, D.-H. Ko, H. Kim, and S.-G. Lee, Design of body gesture recognition system for regularity and repeatability gestures, 14th International Conference on Control, Automation and Systems, pp. 449453, 2014.

[7] J. Fasola and M. J. Mataric, Socially assistive robot exercise coach: Motivating old adults to engage in physical exercise, The 13th International Symposium on Experimental Robotics, pp. 463479, 2013.