A Neural Network Approach for Human Gesture Recognition with a Kinect Sensor

T. D'Orazio[1], C. Attolico[2], G. Cicirelli[1] and C. Guaragnella[2]
[1]*ISSIA - CNR, via Amendola 122/D, Bari, Italy*
[2] *DEI, Politecnico di Bari, Bari, Italy*

Keywords: Feature Extraction, Human Gesture Modelling, Gesture Recognition.

Abstract: Service robots are expected to be used in many household in the near future, provided that proper interfaces are developed for the human robot interaction. Gesture recognition has been recognized as a natural way for the communication especially for elder or impaired people. With the developments of new technologies and the large availability of inexpensive depth sensors, real time gesture recognition has been faced by using depth information and avoiding the limitations due to complex background and lighting situations. In this paper the Kinect Depth Camera, and the OpenNI framework have been used to obtain real time tracking of human skeleton. Then, robust and significant features have been selected to get rid of unrelated features and decrease the computational costs. These features are fed to a set of Neural Network Classifiers that recognize ten different gestures. Several experiments demonstrate that the proposed method works effectively. Real time tests prove the robustness of the method for realization of human robot interfaces.

1 INTRODUCTION

Service robots are expected to be used into every household in the near future. The interaction with keyboard or button is not natural for disabled or elder people. Gesture recognition can be used as an effective and natural way to control robot navigation. Recognition of human gesture from video sequences is a popular task in the computer vision community since it has wide applications including, among others, video surveillance and monitoring, human computer interface, augmented reality, and so on. The use of video sequences of color images made this one a challenging problem due to the interpretation of complex situations in real-life scenarios such as cluttered background, occlusion, illumination and scale variations (Leo et al., 2005; Castiello et al., 2005). The recent availability of depth sensors has allowed the recognition of 3D gesture avoiding a number of problems due to cluttered background, multiple people in the scene, skin color segmentation and so on. In particular, inexpensive Kinect sensors have been largely used by the scientific community as they provide an RGB image and a depth of each pixel in the scene. These sensors have an RGB camera and an infrared (IR) emitter associated with an additional camera, and acquire a new structure of data (RGBD images) that has given rise to a new way to process images. Many functionalities to process sensory data are available in open source frameworks such as OpenNI, and allow the achievement of complex tasks such as people segmentation, real time tracking of a human skeleton (being this one a structure widely used for gesture recognition), scene information, and so on (Cruz et al., 2012).

In the last years many papers have been presented in literature which use the Kinect for human gesture recognition applied in several contexts. In (Cheng et al., 2012; Almetwally and Mallem, 2013) new techniques to imitate the human body motion on humanoid robots were developed using the available skeletal points received from the Kinect Sensors. The ability of the OpenNI framework to easily provide the position and segmentation of the hand has stimulated many approaches on the recognition of hand gestures (den Bergh et al., 2011; J.Oh et al., 2013). The hand orientation and four hand gestures (open, fist,..) are used in (den Bergh et al., 2011) for a gesture recognition system integrated on an interactive robot which looks for a person to interact with, ask for directions, and detects a 3D pointing direction. The motion profiles obtained from the Kinect depth data are used in (Biswas and Basu, 2011) to recognize different gestures by a multi class SVM. The motion information

D'Orazio T., Attolico C., Cicirelli G. and Guaragnella C..
A Neural Network Approach for Human Gesture Recognition with a Kinect Sensor.
DOI: 10.5220/0004919307410746
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 741-746
ISBN: 978-989-758-018-5

is extracted by noting the variation in depth between each pair of consecutive frames. In (Bhattacharya et al., 2012) seven upper body joints are considered for the recognition with SVM of aircraft marshaling gestures used in the military air force. The method requires the data stream editing by a human observer who marks the starting and ending frame of each gesture. The nodes of the skeleton, in (Miranda et al., 2012), are converted in a joint angle representation that provides invariance to sensor orientation. Then a multiclass SVM is used to classify key poses which are forwarded as a sequence to a decision forest for the final gesture recognition. Also in (Gu et al., 2012) joint angles are considered for the recognition of six different gestures but different HMMs have been used to model the gestures. The HMM which provides the maximum likelihood gives the type of gesture. Four joint coordinates relative to the left and right hands and elbows are considered in (Lai et al., 2012) and the normalized distances among these joints form the feature vector which is used in a nearest neighbor classification approach. A rule based approach is used in (Hachaj and Ogiela, 2013) to recognize key postures and body gestures by using an intuitive reasoning module which performs forward chaining reasoning (like a classic expert system) with its inference engine every time new portion of data arrives from the feature extraction library.

In this paper a real time gesture recognition system for human robot interface is presented. We propose a method for gesture recognition from captured skeletons in real time. The experiments were performed using the Kinect platform and the OpenNI framework, in particular considering the positions of the skeleton nodes which are estimated at 30 frames per second. The variations of some joint angles are used as input to several neural network classifiers each one trained to recognize a single gesture. The recognition rate is vey high as will be shown in the experimental section. The rest of the paper is organized as follows: section 2 describes the proposed system, section 3 resumes the results obtained during off-line and on-line experiments. Finally section 4 reports a discussion and conclusions.

2 THE PROPOSED SYSTEM

In this work we propose a Gesture Recognition approach which can be used by a human operator as a natural human computer or human robot interface. In figure 1 both the segmentation and the skeleton provided by the OpenNI framework are shown.We have used the abilities of the Kinect software to identify

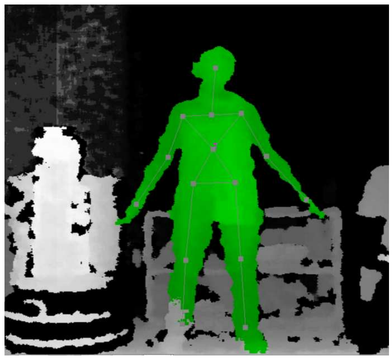

Figure 1: The people segmentation and the skeleton obtained by the Kinect sensor.

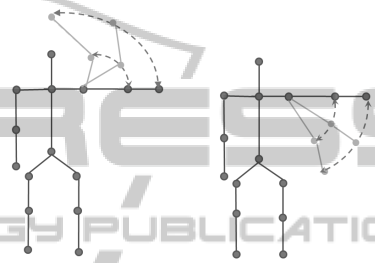

Figure 2: The description of the gestures Attention (G_3) and Mount (G_7).

and track people in the environment and to extract the skeleton with the joint coordinates for the gesture recognition. Ten different gestures executed with the right arm have been selected from the "Arm-and-Hand Signals for Ground Forces" and correspond to the signals *Action Right, Advance, Attention, Contact Right, Fire, Increase Speed, Mount, Move Forward, Move Over, Start Engine*. For the sake of brevity we will refer to these gestures with the abbreviations G_1, G_2, .. G_{10} and we will demand to (SpecialOPeration, 2013) for a detailed description of the selected signals (see figure 2 for the description of G_3 and G_7). The gestures have been executed several times by different persons and some of these executions have been considered for the model generation while the remaining for the test phase.

2.1 Selection of Significant and Robust Features

The problem of selecting significant features which preserve relevant information to the classification, is fundamental for the ability of the method to recognize gestures. Many papers in literature consider the coordinate variations of some joints such as the hand, the elbow, the shoulder and the torso nodes. However when coordinates are considered it is necessary to introduce a kind of normalization in order to be independent of the position and the height of the person in the scene. An alternative way could be the use of angles among joint nodes, but the angle representation is not enough to describe rotation in 3D as the axis of rotation has to be specified to represent a 3D rotation. For this reason we have considered the information provided by the 2.2 OpenNi version, i.e. the quaternions of the joint nodes. The Quaternion is a set of numbers that comprises a four-dimensional vector space and is denoted by $q = a + bi + cj + dk$, where a, b, c, d are real numbers and i, j, k are imaginary units. The quaternion q represents an easy way to code any 3D rotation expressed as a combination of a rotation angle and a rotation axis. In fact starting from the vector q it is possible to extract the rotation angle ϕ and the rotation axis $[v_x, v_y, v_z]$ as follows:

$$\phi = 2 \cdot \arccos a \tag{1}$$

$$[v_x, v_y, v_z] = \frac{1}{\sin(\frac{1}{2}\phi)}[b, c, d] \tag{2}$$

Quaternions can offer fundamental computational implementation and data handling advantages over the conventional rotation matrix. In the considered case, the quaternions of the joint nodes store the direction the bone is pointing to. For the recognition of the ten considered gestures, the quaternions of the shoulder and elbow right nodes have been selected producing a feature vector $V_i = [a_i^s, b_i^s, c_i^s, d_i^s, a_i^e, b_i^e, c_i^e, d_i^e]$, where i is the frame number, the index ^{s} stands for shoulder and ^{e} stands for elbow.

Experimental evidence has demonstrated that, even if the joint position stability is not always guaranteed as it depends on the lighting condition, the reflectivity properties of dresses and so on, the features extracted are robust enough to characterize the periodicity of the gestures. As the user moves freely to give the gesture command, the length of the gesture can be different when executed by different persons. In order to extract more invariant features from the input gesture, the considered features are re-sampled with the same interval and the missing values are interpolated. For more precise recognition result, the total length of each gesture is normalized to 60 frames corresponding to the gesture duration of two seconds.

2.2 Gesture Recognition Algorithm

The models for the gesture recognition have been constructed by using ten different Neural Networks (NN), one for each gesture, which have been trained providing a set of feature sequences of the same gesture as positive examples and the remaining sequences of other gestures as negative examples. Each NN has an input layer of 480 nodes corresponding to the feature vectors V_i for 60 consecutive frames, an hidden layer of 100 nodes and an output layer of one node trained to produce 1 if the gesture is recognized and zero otherwise. The Backpropagation Learning algorithm has been applied and the best configuration of hidden nodes has been selected in an heuristic way after several experiments. At the end of the learning phase, in order to recognize a gesture a sequence of features is provided to all the 10 NNs and the one which returns the maximum value is considered the winning gesture. This classification procedure gives a result also when a gesture does not belong to any of the ten classes. For this reason a threshold has been introduced in order to decide if the maximum answer among the NN outputs has to be assigned to the corresponding class or not.

G	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	**23**	0	0	0	0	0	0	0	0	0
G_2	0	**19**	0	0	0	0	0	0	0	0
G_3	0	0	**16**	0	0	0	0	0	0	0
G_4	0	0	0	**21**	0	0	0	0	0	1
G_5	0	0	0	0	**18**	0	0	0	0	0
G_6	0	0	0	0	0	**15**	0	0	0	0
G_7	0	0	0	0	0	0	**19**	0	0	0
G_8	0	0	0	0	0	0	0	**19**	0	0
G_9	0	0	0	0	0	0	0	0	**21**	0
G_{10}	0	0	0	0	0	0	0	0	0	**18**

Figure 3: The scatter matrix for the recognition of the 10 gestures with *quaternions*. Tests were executed on the same persons used in the training set.

3 EXPERIMENTAL RESULTS

Two different sets of experiments were carried out: off line experiments and on line experiments. In the first case the recognition system has been tested on different persons, extracting manually the sequences of features and comparing the performances with different feature selections. The on line experiments instead were carried out to establish the ability of the proposed algorithm to identify the gesture when different repetitions are proposed and when the system has no knowledge about the starting frame.

3.1 Off-line Experiments

The proposed algorithm was tested using a database of 10 gesture performed by 10 different persons. Selected sequences of gestures performed by five of these persons were used to train the NNs, while the remaining ones together with the sequences of gestures performed by the remaining five persons were used for the test. We will distinguish initially the experiments on the same persons used in the training sets (the first 5 persons), with the experiments executed on the other five persons. We distinguish between these two sets as the execution of the same actions can be very different when they are recorded in different sessions by people who have not seen previous acquisitions. In order to demonstrate that the selected features are representative for the gesture recognition process, first of all a comparative experiment has been carried out considering two types of input features: 1) two quaternions relative to the elbow and the shoulder and 2) three joint-anlges: hand-elbow-shoulder, elbow-shoulder-torso and elbow-rightshoulder-leftshoulder. Then, ten NNs were trained by using sequences of $24, 26, 18, 22, 20, 20, 21, 23, 23, 25$ repetitions of gestures performed by the first 5 persons. Notice that each sequence refers to a gesture: in the order G1, G2, G3 and so on. Each NN used the corresponding set of gestures as positive examples and all the remaining as negative examples. Analogously, the tests were carried out by selecting different executions of the same gestures for a total of $23, 19, 16, 22, 18, 15, 19, 19, 21, 18$ respectively, performed by the same 5 persons. In figures 3 and 4 the scatter matrices of the first two tests are reported: in the first one the results with the *quaternion* and in the second one the results with the *joint angles* are listed. It is evident that the joint angles are not enough to recognize the gestures, while quaternions fail only for one of the sequences of gesture G_4. The experiments with quaternions have been repeated by introducing some sequences which did not belong to any of the ten gestures. In this case in order to recognize a No Gesture class, a threshold was introduced in order to evaluate the maximum value among the NN outputs: if the maximum outputs is under the threshold the gesture is classified as No Gesture. As a consequence some gestures that were correctly classified in the scatter matrix reported in 4 are now considered as No Gesture in 5 since the maximum value is under the threshold of 0.7 (see the last column NG). According to the threshold value the number of false negatives (gesture recognized as No gesture) and true negatives (No gesture detected as No gesture) can greatly change. In figure 6 the experiments have been repeated on sequences acquired by persons that were non included in the training set and using a threshold for the detection of No gesture. As expected the maximum answers of the NNs are smaller than those obtained in the previous experiments, but the recognition of many of the ten gestures is however guaranteed using a smaller threshold value of 0.4.

G	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	**23**	0	0	0	0	0	0	0	0	0
G_2	0	**19**	0	0	0	0	0	0	0	0
G_3	0	0	**16**	0	0	0	0	0	0	0
G_4	0	0	0	**16**	0	0	0	0	0	6
G_5	0	0	0	0	**18**	0	0	0	0	0
G_6	0	0	0	0	0	**14**	0	0	1	0
G_7	0	0	0	0	0	0	**19**	0	0	0
G_8	0	0	0	0	0	0	0	**19**	0	0
G_9	1	0	0	7	0	0	0	0	**13**	0
G_{10}	0	0	0	0	0	0	0	0	0	**18**

Figure 4: The scatter matrix for the recognition of the 10 gestures with *joint angles*.Tests were executed on the same persons used in the training set.

G	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	NG
G_1	**23**	0	0	0	0	0	0	0	0	0	0
G_2	0	**18**	0	0	0	0	0	0	0	0	1
G_3	0	0	**16**	0	0	0	0	0	0	0	0
G_4	0	0	0	**21**	0	0	0	0	0	0	**1**
G_5	0	0	0	0	**18**	0	0	0	0	0	0
G_6	0	0	0	0	0	**15**	0	0	0	0	0
G_7	0	0	0	0	0	0	**18**	0	0	0	1
G_8	0	0	0	0	0	0	0	**19**	0	0	0
G_9	0	0	0	0	0	0	0	0	**21**	0	0
G_{10}	0	0	0	0	0	0	0	0	0	**18**	0
NG	1	1	0	0	0	0	0	1	6	3	**7**

Figure 5: The scatter matrix for the recognition of the 10 gestures with *quaternions* with a threshold of 0.7 on the NN maximum value for the detection of NoGesture.

G	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	NG
G_1	**19**	0	0	0	0	0	0	0	0	0	6
G_2	0	**11**	0	0	0	0	0	0	1	0	1
G_3	0	0	**12**	0	0	0	0	0	0	0	3
G_4	0	0	0	**20**	0	0	0	0	0	0	**5**
G_5	0	0	0	0	**8**	0	0	0	0	0	12
G_6	0	0	0	0	0	**18**	0	0	0	0	0
G_7	0	0	0	0	0	0	**16**	0	0	0	2
G_8	0	0	0	0	0	0	0	**18**	0	0	6
G_9	0	0	0	0	0	0	0	0	**26**	0	0
G_{10}	0	0	0	0	0	0	0	0	0	**18**	0
NG	1	1	0	0	0	0	0	1	6	3	**7**

Figure 6: The same experiment of figure 5 on different persons not included in the training set. The threshold for the detection of NoGesture is 0.4.

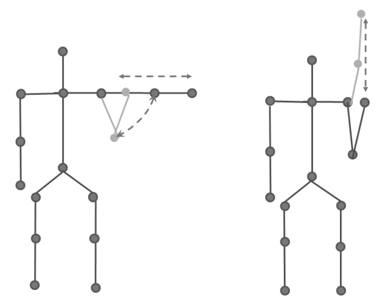

Figure 7: The description of the gestures Action Right (G_1) and Increase Speed (G_6).

3.2 On Line Experiments

In the off line experiments the sequences were manually extracted both for the training and the test of the ten NNs. However for the real application of such a recognition system we have to consider that it is not possible to know the starting frame of each gesture. For this reason another set of experiments, named on line experiments, was carried out. Different persons were asked to repeat the same gesture without interruption and all the frames of the sequences were processed. A sliding window of 60 frames was extracted and fed to all the 10 NNs. In figure 8 the output of the NNs for the execution of the Gesture 1 over a sequence of 600 frames, are reported. In red (○) the correct answers of the NN1 and in green (+) the wrong answers of the NN6 are shown. It is evident that for most of the time the maximum answers are provided by the correct NN1 whereas in some regular intervals the NN6 provides larger values. The wrong answers are justified by the fact that if the sliding windows is not centered at the beginning of the gestures then the NN1 provides values lower than NN6. In particular observing these two gestures (G_1 and G_6), the central part of the $G1$ is very similar to the beginning of G_6 (see figure 7). A filter on the number of consecutive concordant answers is applied (35 frames in this case). The results are reported in figure 9: for most of the time G_1 is correctly recognized while in the remaining intervals a no gesture is associated. The same conclusions are confirmed by the results reported in figures 10 and 11 where G_7 has been repeated by another person several times. The gesture is in part confused with G_9 but with the filter on the number of concordant answers only G_7 is correctly recognized.

4 DISCUSSION AND CONCLUSIONS

In this paper we have proposed a gesture recognition system based on a Kinect sensor, a low cost RGBD camera, which provides people segmentation in an effective way and skeleton information for real time processing. We used the quaternion features of the right shoulder and elbow nodes to construct the models of 10 different gestures. Thus ten different Neural Networks have been trained using a set of positive examples of the corresponding correct gesture and the remaining gestures as negative examples. Off-line experiments have demonstrated that the NNs are able to model the considered gestures when the sequences of frames corresponding to the gestures of both persons included and not included in the training set, are manually extracted during a set of acquisitions. However the knowledge of the initial frame in which the gesture starts is not always guaranteed during on line experiments, for this reason we have imposed the constraint on the repetition of the same action for a number of consecutive times, in order to make a decision by observing the results of a sliding window moved

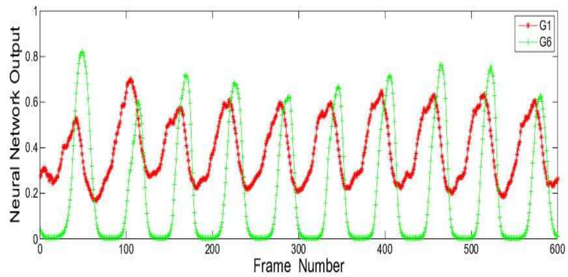

Figure 8: The results of the gesture recognition over a sequence of 600 frames during which G_1 has been executed in a continuous way.

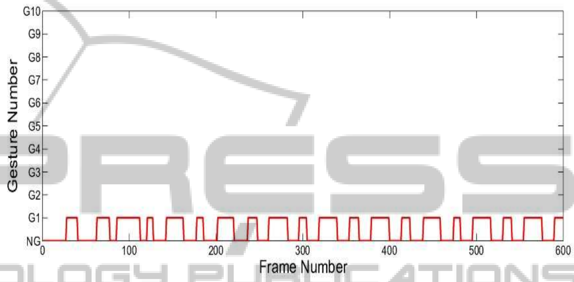

Figure 9: The results of the gesture recognition when a threshold on the number of consecutive answers is applied: G_1 is correctly recognized.

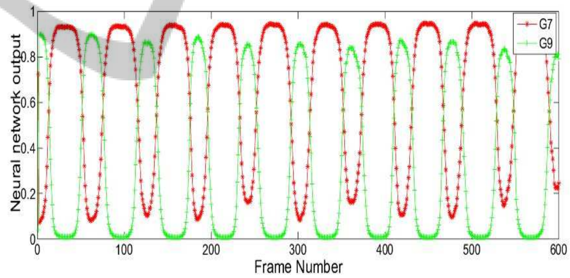

Figure 10: The results of the gesture recognition over a sequence of 600 frames during which G_7 has been executed in a continuous way.

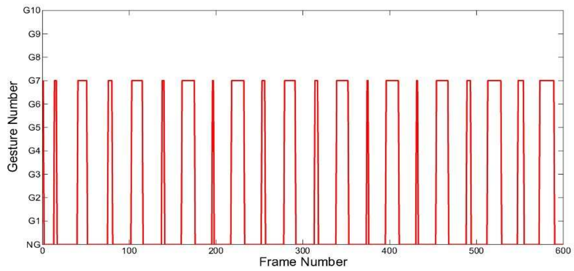

Figure 11: The results of the gesture recognition when a threshold on the number of consecutive answers is applied: $G7$ is correctly recognized.

over the entire sequence. In this way we are sure that when the sliding window is centered around the gesture the corresponding NN will provide the maximum answers, while when the window overlaps the ending or the beginning parts of the gestures some false positive answers can be provided. The obtained results are very encouraging as the number of false positives is always smaller than the true positives. Furthermore by filtering on the number of consecutive concordant answers a correct final decision can be taken. Tests executed on persons different from those used in the training set have demonstrated that the proposed system can be trained off line and used for the gesture recognition by any other user with the only constraint of repeating the same gesture more times.

In future work we will face the problem of the length of the gestures. In this paper we have imposed that the gestures are all executed in 2 seconds corresponding to 60 frames. When the gestures are executed with different velocities the correct association is not guaranteed. Current researches focus on the automatic detection of the gesture length and on the normalization of all the executions by interpolating the missing values.

ACKNOWLEDGEMENTS

This research has been developed under grant PON 01-00980 BAITAH.

REFERENCES

Almetwally, I. and Mallem, M. (2013). Real-time teleoperation and tele-walking of humanoid robot nao using kinect depth camera. *10th IEEE International Conference on Networking, Sensing and Control (ICNSC)*, page 463466.

Bhattacharya, S., Czejdo, B., and Perez, N. (2012). Gesture classification with machine learning using kinect sensor data. *Third International Conference on Emerging Applications of Information Technology (EAIT)*, pages 348 – 351.

Biswas, K. and Basu, S. (2011). Gesture recognition using microsoft kinect. *5th International Conference on Automation, Robotics and Applications (ICARA)*, pages 100–103.

Castiello, C., D'Orazio, T., Fanelli, A., Spagnolo, P., and Torsello, M. (2005). A model free approach for posture classificatin. *IEEE Conf. on Advances Video and Signal Based Surveillance, AVSS*.

Cheng, L., Sun, Q., Cong, Y., and Zhao, S. (2012). Design and implementation of human-robot interactive demonstration system based on kinect. *24th Chinese Control and Decision Conference (CCDC)*, page 971975.

Cruz, L., Lucio, F., and Velho, L. (2012). Kinect and rgbd images: Challenges and applications. *XXV SIBGRAPI IEEE Confernce and Graphics, Patterns and Image Tutorials*, page 3649.

den Bergh, M. V., Carton, D., de Nijs, R., Mitsou, N., Landsiedel, C., Kuehnlenz, K., Wollherr, D., Gool, L. V., and Buss, M. (2011). Real-time 3d hand gesture interaction with a robot for understanding directions from humans. *20th IEEE international symposium on robot and human interactive communication*, pages 357 – 362.

Gu, Y., andY. Ou, H. D., and Sheng, W. (2012). Human gesture recognition through a kinect sensor. *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1379 – 1384.

Hachaj, T. and Ogiela, M. (2013). Rule-based approach to recognizing human body poses and gestures in real time. *Multimedia Systems*.

J.Oh, Kim, T., and Hong, H. (2013). Using binary decision tree and multiclass svm for human gesture recognition. *International Conference on Information Science and Applications (ICISA)*, pages 1 – 4.

Lai, K., Konrad, J., and Ishwar, P. (2012). A gesture-driven computer interface using kinect. *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 185 – 188.

Leo, M., P.Spagnolo, D'Orazio, T., and Distante, A. (2005). Human activity recognition in archaeological sites by hidden markov models. *Advances in Multimedia Information Procesing - PCM 2004*.

Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A., and Campos, M. (2012). Real-time gesture recognition from depth data through key poses learning and decision forests. *25th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 268 – 275.

SpecialOPeration (2013). Arm-and-hand signals for ground forces. *www.specialoperations.com/Focus/Tactics/Hand_Signals/default.htm*.