Vision-based Gesture Recognition System for Human-Computer Interaction

Paulo Trigueiros
IPP-Instituto Politécnico do Porto, Porto, Portugal
ptrigueiros@gmail.com
Fernando Ribeiro
EEUM-Escola de Engenharia da Universidade do Minho - DEI, Guimarães, Portugal
fernando@dei.uminho.pt
Luís Paulo Reis
EEUM-Escola de Engenharia da Universidade do Minho – DSI, Guimarães, Portugal
lpreis@dsi.uminho.pt

ABSTRACT: Hand gesture recognition, being a natural way of human computer interaction, is an area of active research in computer vision and machine learning. This is an area with many different possible applications, giving users a simpler and more natural way to communicate with robots/systems interfaces, without the need for extra devices. So, the primary goal of gesture recognition research is to create systems, which can identify specific human gestures and use them to convey information or for device control. This work intends to study and implement a solution, generic enough, able to interpret user commands, composed of a set of dynamic and static gestures, and use those solutions to build an application able to work in a real-time human-computer interaction systems. The proposed solution is composed of two modules controlled by a FSM (Finite State Machine): a real time hand tracking and feature extraction system, supported by a SVM (Support Vector Machine) model for static hand posture classification and a set of HMMs (Hidden Markov Models) for dynamic single stroke hand gesture recognition. The experimental results showed that the system works very reliably, being able to recognize the set of defined commands in real-time. The SVM model for hand posture classification, trained with the selected hand features, achieved an accuracy of 99,2%. The proposed solution as the advantage of being computationally simple to train and use, and at the same time generic enough, allowing its application in any robot/system command interface.

Keywords: Human-Computer Interaction, Gesture Recognition, SVM, ANN, HMM, FSM

1 INTRODUCTION

Hand gesture recognition for human computer interaction is an area of active research in computer vision and machine learning. The primary goal of gesture recognition research, is to create a system, which can identify specific gestures and use them to convey information or for device control. For that, gestures need to be modelled in the spatial and temporal domains, where a hand posture is the static structure of the hand and a gesture is the dynamic movement of the hand. Being hand-pose one of the most important communication tools in human's daily life, and with the continuous advances of image and video processing techniques, research on human-machine interaction through gesture recognition led to the use of such technology in a very broad range of applications, like touch screens, video game consoles, virtual reality, medical applications, etc. There are areas where this trend is an asset, as for example in the application of these technologies in interfaces that can help people with physical disabilities, or areas where it is a complement to the normal way of communicating.

There are basically two types of approaches for hand gesture recognition: vision-based approaches and data glove methods. In the study we will be focusing our attention on vision-based approaches. Why vision-based hand gesture recognition systems? Vision-based hand gesture recognition systems provide a simpler and more intuitive way of communication between a human and a computer. Using visual input in this context makes it possible to communicate remotely with computerized equipment, without the need for physical contact. The main objective of this work is to study and implement solutions that can be generic enough, with the help of machine learning algorithms, allowing its application in a wide range of human-computer interfaces, for online gesture recognition. In pursuit of this, we intend to use a depth sensor camera to detect and extract hand information (hand features), for gesture classification. With the implemented solutions we intend to develop an integrated vision-based hand gesture recognition system, for offline training of static and dynamic hand gestures, in order to create

models, that can be used for online classification of user commands.

The rest of the paper is as follows. Firstly, the related work is review in section 2. Section 3 introduces the system architecture and describes hand posture classification and dynamic gesture classification. Experimental methodology and results are explained in section 4. Conclusions and further work are drawn in section 5.

2 RELATED WORK

Hand gestures, either static or dynamic, for human computer interaction in real time systems is an area of active research and with many possible applications. However, vision-based hand gesture interfaces for real-time applications require fast and extremely robust hand detection, feature extraction and gesture recognition. Several approaches are normally used including Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Hidden Markov Models (HMM).

An Artificial Neural Networks is a mathematical / computational model that attempts to simulate the structure of biological neural systems. They accept features as inputs and produce decisions as outputs (Snyder and Qi, 2004). Maung et al (Maung, 2009) applied it in a gesture recognition system for real-time gestures in unstrained environments. Vicen-Buéno et al. (Vicen-Bueno et al., 2004) used it applied to the problem of traffic sign recognition. Bailador et al. (Bailador et al., 2007) presented an approach to the problem of gesture recognition in real time using inexpensive accelerometers. Their approach was based on the idea of creating specialized signal predictors for each gesture class.

A Support Vector Machines (SVM's) is a technique based on statistical learning theory, which works very well with high-dimensional data. The objective of this algorithm is to find the optimal separating hyper plane between two classes by maximizing the margin between them (Ben-Hur and Weston, 2008). Faria et al. (Faria et al., 2009, Faria et al., 2010) used it to classify robotic soccer formations and the classification of facial expressions, Ke et al. (Ke et al., 2010) used it in the implementation of a real-time hand gesture recognition system for human robot interaction, Maldonado-Báscon (Maldonado-Báscon et al., 2007) used it for the recognition of road-signs and Masaki et al. (Oshita and Matsunaga, 2010) used it in conjunction with SOM (Self-Organizing Map) for the automatic learning of a gesture recognition mode. He first applies the SOM to divide the sample into phases and construct a state machine, and then he applies the SVM to learn the transition conditions between nodes. Almeida et al. (Almeida et al., 2009) proposed a classification approach to identify the team's formation in the robotic soccer domain for the two dimensional (2D) simulation league employing Data Mining classification techniques. Trigueiros et al. (Trigueiros et al., 2012) have made a comparative study of four machine learning algorithms applied to two hand features datasets. In their study the datasets had a mixture of hand features.

Hidden Markov Models (HMMs) have been widely used in a successfully way in speech recognition and hand writing recognition (Rabiner, 1989), in various fields of engineering and also applied quite successfully to gesture recognition. Oka et al. (Oka et al., 2002) developed a gesture recognition system based on measured finger trajectories for an augmented desk interface system. They have used a Kalman filter for the prediction of multiple finger locations and used an HMM for gesture recognition. Perrin et al. (Perrin et al., 2004) described a finger tracking gesture recognition system based on a laser tracking mechanism which can be used in hand-held devices. They have used HMM for their gesture recognition system with an accuracy of 95% for a set of 5 gestures. Nguyen et al. (Binh et al., 2005) described a hand gesture recognition system using a real-time tracking method with pseudo two-dimensional Hidden Markov Models. Chen et al. (Chen et al., 2003) used it in combination with Fourier descriptors for hand gesture recognition using a real-time tracking method. Kelly et al. (Kelly et al., 2011) implemented an extension to the standard HMM model to develop a gesture threshold HMM (GT-HMM) framework which is specifically designed to identify inter gesture transition. Zafrulla et al. (Zafrulla et al., 2011) have investigated the potential of the Kinect depth-mapping camera for sign language recognition and verification for educational games for deaf children. They used 4-state HMMs to train each of the 19 signs defined in their study. Cooper et al. (Helen and Richard, 2007) implemented an isolated sign recognition system using a 1st order Markov chain. In their model, signs are broken down in visemes (equivalent to phonemes in speech) and a bank of Markov chains are used to recognize the visemes as they are produced. Milosevic et al. (Milosevic et al., 2010) implemented an HMM-based continuous gesture recognition algorithm, optimized for lower-power, low-cost microcontrollers without float point unit. The proposed solution is validated on a set of gestures performed with the Smart Micrel Cube (SMCube), which embeds a 3-axis accelerometer and an 8-bit microcontroller. They also explore a multiuser scenario where up to 4 people share the same device. Elmezain et al. (Elmezain et al., 2008) proposed a system able to recognize both isolated and continuous gestures for Arabic numbers (0-9)

in real-time. To handle isolated gestures, an HMM using Ergodic (it is possible to go from every state to every state), Left-Right (LR) and Left-Right Banded (LRB) topologies with different number of states was applied. The LRB in conjunction with the Forward algorithm presented the best performance with an average recognition rate of 98,94% and 95,7% for isolated and continuous gestures.

3 PROPOSED SYSTEM ARCHITECTURE

The design of any gesture recognition system essentially involves the following three aspects: (1) *data acquisition and pre-processing*; (2) *data representation or feature extraction* and (3) *classification or decision-making*. Taking this into account, a possible solution to be used in any vision-based hand gesture recognition system for human-computer interaction is represented in the following diagram.

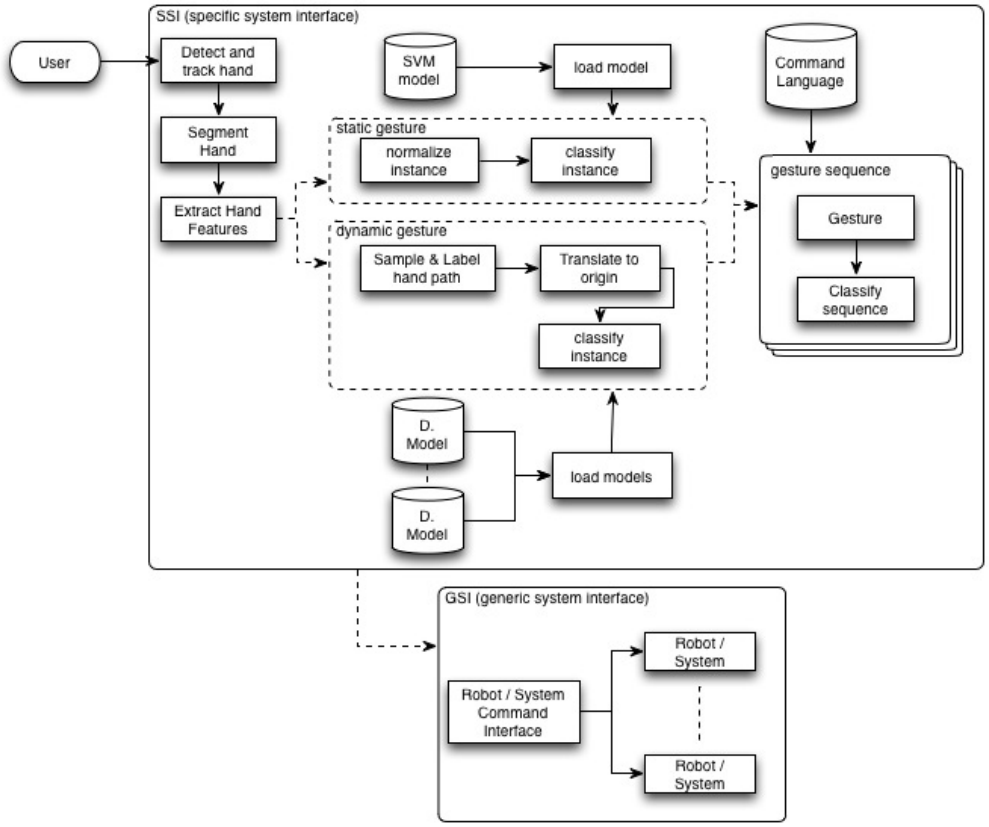

Figure 1. Vision-based hand gesture recognition system architecture.

In the following sections, we will describe the problems of hand posture classification , dynamic gesture classification and gesture sequence or command classification.

3.1 Hand posture classification

For static gesture classification, hand segmentation and feature extraction is a crucial step in vision-based hand gesture recognition systems. The pre-processing stage prepares the input image and extracts features used later with classification algorithms (Trigueiros et al., 2013). Our system uses feature vectors composed of centroid distance values to train a SVM (Support Vector Machine) for hand posture classification. The centroid distance signature is a type of shape signature (Trigueiros et al., 2013) expressed by the distance of the hand contour boundary points, from the hand centroid (x_c, y_c) and is calculated in the following manner:

$$d(i) = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}, i = 0, \ldots, N-1 \quad (1)$$

This way, a one-dimensional function representing the hand shape is obtained. The number of equally spaced points N used in our implementation was 32.

The SVM is used to learn the pre-set hand postures shown in Figure 2. The SVM is a pattern recognition technique in the area of supervised machine learning, which works very well with high-dimensional data. When more than two classes are present, there are several approaches that evolve around the 2-class case (Theodoridis and Koutroumbas, 2010). The one used in this system is the one-against-all, where *c* classifiers have to be designed. Each one of them is designed to separate one class from the rest.

Figure 2. The defined hand postures for the static commands.

3.2 Dynamic gesture classification

Dynamic gestures are time-varying processes, which show statistical variations, making HMMs a plausible choice for modelling the processes (Rabiner and Juang, 1986) (Wu and Huang, 1999). So, for the recognition of dynamic gestures a HMM (Hidden Markov Model) model was trained for each possible gesture. A Markov Model is a typical model for a stochastic (i.e. random) sequence of a finite number of states (Fink, 2008). A human gesture can be understood as a Hidden Markov Model where the true states of the model are hidden in the sense that they cannot be directly observed. HMMs have been widely used in a successfully way in speech recognition and hand writing recognition (Rabiner, 1989). In this system the 2D motion hand trajectory points are labelled according to the distance to the nearest centroid, based on Euclidean distance, and translated to origin resulting in a discrete feature vector like the one shown in Figure 3. In the proposed solution, the 2D features are sufficient for hand gesture recognition.

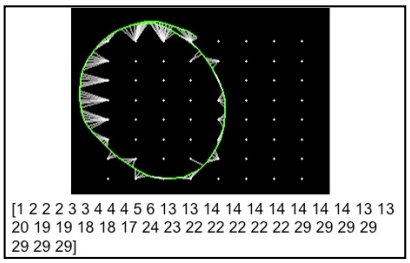

Figure 3. Gesture path with respective feature vector.

The feature vectors thus obtained are used to train the different HMMs and learn the model parameters: the initial state probability vector (π), the state-transition probability matrix ($A = [a_{ij}]$) and the observable symbol probability matrix ($B = [b_j(m)]$). In the recognition phase an output score for the sample gesture is calculated for each model, given the likelihood that the corresponding model generated the underlying gesture. The model with the highest output score represents the recognized gesture. In our system a Left-Right (LR) HMM, like the one shown in Figure 4, was used (Camastra and Vinciarelli, 2008, Alpaydin, 2004). This kind of HMM has the states ordered in time so that as time increases, the state index increases or stays the same. This topology has been chosen, since it is perfectly suitable to model the kind of temporal gestures present in the system.

3.3 Command classification

Since the system uses a combination of dynamic and static gestures, modelling the command semantics became necessary. A Finite State Machine is a usually employed technique to handle this situation (Buckland, 2005, Millington and Funge, 2009). This system uses a FSM, as shown in Figure 5, to control the transition between three possible states: *DYNAMIC*, *STATIC* and *PAUSE*. By using a pause state it is possible to identify transitions between gestures and somehow eliminate all unintentional actions between dynamic/static or static/static gestures. The combination of a dynamic gesture and a static or set of static gestures gives us a user command that can then be transmitted to a Generic System Interface (GSI) shown in Figure 1.

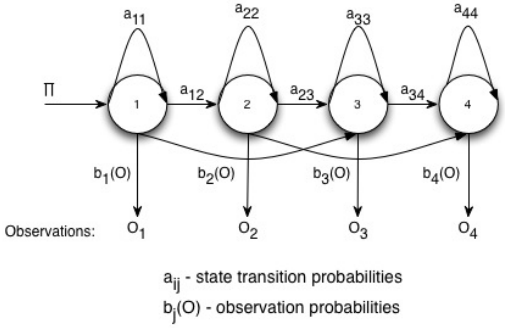

Figure 4. A 4-state Left-Right HMM model.

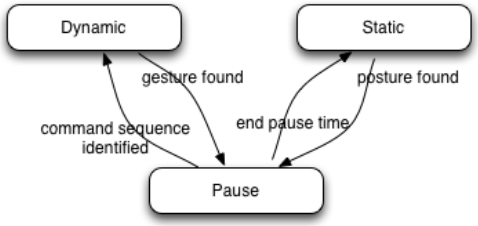

Figure 5. The Finite State Machine (FSM) diagram.

4 EXPERIMENTAL METHODOLOGY AND RESULTS

The experimental methodology was divided into three parts: a feature extraction phase and SVM model training for the defined set of hand postures; an hand motion sequence acquisition phase for each of the defined gestures and HMM model training; and the Vision-based prototype implementation, which is a system with full gesture recognition, controlled by a event driven Finite State Machine that controls the transition between the three possible states (*DYNAMIC*, *STATIC* and *PAUSE*), and builds the final command. For hand posture recognition an SVM model was trained based on a dataset build from data collected from four users making the defined postures in front of a Kinect camera. In order to

analyse the best parameters for SVM, the extracted features were analysed with the help of Rapid Miner (Miner). The experiments were performed with parameter optimization for the C and gamma values, with a 10-fold cross validation, and we obtained an accuracy of 99,2% with an RBF (radial basis function) kernel with a C value equal to 2 and a gamma value of 0.1. The obtained confusion matrix can be seen in **Table 1**, where it is possible to see the existence of some high values, marked red, between command number four and three, between command number five and four and between command number six and seven, which contributed to the 0,8% of false positives. For model training and implementation we used Dlib (King, 2009), a general-purpose cross-platform C++ library capable of SVM multiclass classification.

Table 1. Centroid distance confusion matrix

		Actual class						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Predicted class	**1**	**602**	0	0	0	0	0	0
	2	2	**712**	0	0	1	0	1
	3	0	1	**578**	1	0	0	0
	4	0	0	12	**715**	3	0	0
	5	0	1	1	13	**542**	1	3
	6	1	2	0	1	5	**701**	12
	7	0	0	0	2	0	1	**751**

For dynamic gesture recognition, an HMM model was trained off-line for each one of the 11 predefined gestures. Once again four users were used to perform the predefined gestures and the extracted features were saved and used to train the models. For the HMM models, the number of centroids (alphabet) for gesture quantization, is a predefined value, equally spaced in a 2D grid as explained in section 3.2. For each gesture an HMM was trained and the three parameters (*the initial state probability vector, the state-transition probability matrix and the observable symbol probability matrix*) were learned and saved. The number of observation symbols (alphabet) and hidden states were learned by trial and error, and were defined to be 64 and 4 respectively. For values of observation symbols superior to 64 no significant improvement was noticed. The online recognition tests proved that the created models were capable of reliably identify all the trained gestures. For model training and implementation it was decided to use an openFrameworks (openframeworks.cc) add-on implementation of the HMM algorithm for classification and recognition of numeric sequences. This implementation is a C++ porting of a MATLAB code from Kevin Murphy (Murphy, 1998).

5 CONCLUSIONS AND FUTURE WORK

This paper presented a system able to interpret dynamic and static gestures from a user with the goal of real-time human-computer interaction. Although the machine learning algorithms used are not the only solutions, they were selected based on obtained performance, and the type of system that we wanted to implement. Thus, for static gesture recognition an SVM model was learned from centroid distance features and a recognition rate of (99,2%) was achieved. SVM's are not the only solution, but the experiments showed that this machine learning algorithm achieved the best results with the features under study. For temporal gesture recognition, as stated in section 3.2, HMMs are a plausible choice for dynamic gesture recognition, so one HMM was trained for each gesture. We were able to test the system in real time situations, and it was possible to prove from the experiments that the trained models were able to recognize all the trained gestures, proving that this kind of models, as was already seen in other references, works very well for this type of problem. The experimental results also showed, that the proposed system was able to reliably recognize the pre-defined commands (combination of dynamic and static gestures) in real-time, although with some limitations imposed by the presence of image noise and the low frequency frame rates achieved, due to the nature of the camera used.

The system only uses 2D gesture paths for dynamic gestures, although as future work it is our intention to test and include not only the possibility of 3D dynamic gestures but also to work with several cameras to thereby obtain a full 3D environment and achieve view-independent recognition. Also, the existence of different new cameras, with improved depth resolution and with a higher frame frequency, now available on the market will be tested, to improve some of the limitations of the Kinect, thus leading to better static and dynamic gesture recognition rates.

6 REFERENCES

ALMEIDA, R., REIS, L. P. & JORGE, A. M. 2009. Analysis and Forecast of Team Formation in the Simulated Robotic Soccer Domain. *Proceedings of the 14th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence.* Aveiro, Portugal: Springer-Verlag.

ALPAYDIN, E. 2004. Introduction to Machine Learning, MIT Press.

BAILADOR, G., ROGGEN, D., TRÖSTER, G. & TRIVIÑO, G. 2007. Real time gesture recognition using continuous time recurrent neural networks. *Proceedings of the ICST 2nd international conference on Body area*

networks. Florence, Italy: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

BEN-HUR, A. & WESTON, J. 2008. A User's Guide to Support Vector Machines. *Data Mining Techniques for the Life Sciences.* Humana Press.

BINH, N. D., SHUICHI, E. & EJIMA, T. Real-Time Hand Tracking and Gesture Recognition System. Proceedings of International Conference on Graphics, Vision and Image December 2005 2005 Cairo - Egypt. 362--368.

BUCKLAND, M. 2005. *Programming Game AI by Example*, Wordware Publishing, Inc.

CAMASTRA, F. & VINCIARELLI, A. 2008. Machine Learning for Audio, Image and Video Analysis, Springer.

CHEN, F.-S., FU, C.-M. & HUANG, C.-L. 2003. Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing,* 21**,** 745-758.

ELMEZAIN, M., AL-HAMADI, A., APPENRODT, J. & MICHAELIS, B. A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory. Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 8-11 Dec. 2008 2008. 1-4.

FARIA, B. M., LAU, N. & REIS, L. P. Classification of Facial Expressions Using Data Mining and machine Learning Algorithms. *In:* AISTI, ed. 4ª Conferência Ibérica de Sistemas e Tecnologias de Informação, 17 a 20 Junho de 2009 2009 Póvoa de Varim, Portugal. 197-206.

FARIA, B. M., REIS, L. P., LAU, N. & CASTILLO, G. 2010. Machine Learning Algorithms applied to the Classification of Robotic Soccer Formations ans Opponent Teams. *IEEE Conference on Cybernetics and Intelligent Systems (CIS).* Singapore.

FINK, G. A. 2008. Markov Models for Pattern recognition - From Theory to Applications, Springer.

HELEN, C. & RICHARD, B. 2007. Large lexicon detection of sign language. *IEEE International Conference on Human-Computer Interaction.* Rio de Janeiro, Brazil: Springer-Verlag.

KE, W., LI, W., RUIFENG, L. & LIJUN, Z. 2010. Real-Time Hand Gesture Recognition for Service Robot. 976-979.

KELLY, D., MCDONALD, J. & MARKHAM, C. 2011. Recognition of Spatiotemporal Gestures in Sign Language Using Gesture Threshold HMMs. *In:* WANG, L., ZHAO, G., CHENG, L. & PIETIK√§INEN, M. (eds.) *Machine Learning for Vision-Based Motion Analysis.* Springer London.

KING, D. E. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research,* 10**,** 1755-1758.

MALDONADO-BÁSCON, S., LAFUENTE-ARROYO, S., GIL-JIMÉNEZ, P. & GÓMEZ-MORENO, H. 2007. Road-Sign detection and Recognition Based on Support Vector Machines. *IEEE Transactions on Intelligent Transportation Systems.*

MAUNG, T. H. H. 2009. Real-Time Hand Tracking and Gesture Recognition System Using Neural Networks. *Proceedings of World Academy of Science: Engineering & Technology,* 50**,** 466-470.

MILLINGTON, I. & FUNGE, J. 2009. *Artificial Intelligence for Games*, Elsevier.

MILOSEVIC, B., FARELLA, E. & BENINI, L. Continuous Gesture Recognition for Resource Constrained Smart Objects. *In:* IARIA, ed. UBICOMM 2010 : The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, October 25 - 30 2010 Florence, Italy. 391-396.

MINER, R. *RapidMiner : Report the Future* [Online]. Available: http://rapid-i.com/content/view/181/196/ [Accessed December 2011.

MURPHY, K. 1998. *Hidden Markov Model (HMM) Toolbox for Matlab* [Online]. Available: Hidden Markov Model (HMM) Toolbox for Matlab 2013].

OKA, K., SATO, Y. & KOIKE, H. 2002. Real-time fingertip tracking and gesture recognition. *Computer Graphics and Applications, IEEE,* 22**,** 64-71.

OSHITA, M. & MATSUNAGA, T. Automatic learning of gesture recognition model using SOM and SVM. *In:* BEBIS, G., BOYLE, R., PARVIN, B., KORACIN, D. & CHUNG, R., eds. International Conference on Advances in Visual Computing, 2010. Springer-Verlag, 751-759.

PERRIN, S., CASSINELLI, A. & ISHIKAWA, M. Gesture recognition using laser-based tracking system. Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, 17-19 May 2004 2004. 541-546.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* 77**,** 257-286.

RABINER, L. R. & JUANG, B. H. 1986. An introduction to hidden Markov models. *IEEE ASSp Magazine*.

SNYDER, W. E. & QI, H. 2004. *Machine Vision*, Cambridge University Press.

THEODORIDIS, S. & KOUTROUMBAS, K. 2010. *An Introduction to Pattern Recognition: A Matlab Approach*, Academic Press.

TRIGUEIROS, P., RIBEIRO, F. & REIS, L. P. A comparison of machine learning algorithms applied to hand gesture recognition. 7ª Conferência Ibérica de Sistemas e Tecnologias de Informação, 2012 Madrid, Spain. 41-46.

TRIGUEIROS, P., RIBEIRO, F. & REIS, L. P. A Comparative Study of different image features for hand gesture machine learning. 5th International Conference on Agents and Artificial Intelligence, 15-18 February 2013 Barcelona.

VICEN-BUENO, R., GIL-PITA, R., JARABO-AMORES, M. P. & LÓPEZ-FERRERAS, F. 2004. Complexity Reduction in Neural Networks Appplied to Traffic Sign Recognition Tasks.

WU, Y. & HUANG, T. S. 1999. Vision-Based Gesture Recognition: A Review. Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction. Springer-Verlag.

ZAFRULLA, Z., BRASHEAR, H., STARNER, T., HAMILTON, H. & PRESTI, P. 2011. American sign language recognition with the kinect. *Proceedings of the 13th international conference on multimodal interfaces.* Alicante, Spain: ACM.