A Comparative Study of Different Image Features for Hand Gesture Machine Learning

Paulo Trigueiros[1], Fernando Ribeiro[2] and Luís Paulo Reis[3]
[1,2] *Departamento de Electrónica Industrial da Universidade do Minho, Campus de Azurém 4800-05, Guimarães, Portugal*
[3]*EEUM – Escola de Engenharia da Universidade do Minho – DSI, Campus de Azurém 4800-058, Guimarães, Portugal*

Keywords: Hand Gesture Recognition, Machine Vision, Hand Features, Hog, Fourier Descriptors, Centroid Distance, Radial Signature, Shi-Tomasi Corner Detection.

Abstract: Vision-based hand gesture interfaces require fast and extremely robust hand detection, and gesture recognition. 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 human gestures and use them to convey information or for device control. In this paper we present a comparative study of seven different algorithms for hand feature extraction, for static hand gesture classification, analysed with RapidMiner in order to find the best learner. We defined our own gesture vocabulary, with 10 gestures, and we have recorded videos from 20 persons performing the gestures for later processing. Our goal in the present study is to learn features that, isolated, respond better in various situations in human-computer interaction. Results show that the radial signature and the centroid distance are the features that when used separately obtain better results, being at the same time simple in terms of computational complexity.

1 INTRODUCTION

Hand gesture recognition, being a natural way of human computer interaction, is an area of active current research, with many different possible applications, in order to create simpler and more natural forms of interaction, without using extra devices (Hninn and Maung 2009; Trigueiros, Ribeiro et al. 2012).

To achieve natural human-computer interaction, the human hand could be considered as an input device. Hand gestures are a powerful way of human communication, with lots of potential applications, and vision-based hand gesture recognition techniques have many proven advantages compared with traditional devices. Compared with traditional HCI (Human Computer Interaction) devices, hand gestures are less intrusive and more convenient to explore, for example, three-dimensional (3D) virtual worlds. However, the expressiveness of hand gestures has not been fully explored for HCI applications. So, hand gesture recognition has become a challenging topic of research. However, recognizing the shape (posture) and the movement (gesture) of the hand in images or videos is a complex task (Bourennane and Fossati 2010).

The approach used for the problem in vision-based hand gesture recognition consists of identifying the pixels on the image that constitute the hand, extract features from those identified pixels in order to classify the hand, and use those features to recognize the occurrence of specific pose or sequence of poses as gestures.

In this paper we present a comparative study of seven different algorithms for hand feature extraction, for static hand gesture classification. The features were analysed with RapidMiner (http://rapid-i.com) in order to find the best learner, among the following four: k-NN, Naïve Bayes, ANN and SVM. We defined our own gesture vocabulary, with 10 gestures as shown in Figure 1, and we have recorded videos from 20 persons performing the gestures for later processing. Our goal in the present study is to learn features that, isolated, respond better in various situations in human-computer interaction. The results show that the radial signature and the centroid distance are the features that when used separately obtain better results, being at the same time simple in terms of

Trigueiros P., Ribeiro F. and Reis L..
A Comparative Study of Different Image Features for Hand Gesture Machine Learning.
DOI: 10.5220/0004200100510061
In *Proceedings of the 5th International Conference on Agents and Artificial Intelligence* (ICAART-2013), pages 51-61
ISBN: 978-989-8565-39-6

computational complexity. The features were selected due to their computational simplicity and efficiency in terms of computation time, and also because of the good recognition rates shown in other areas of study, like human detection (Dalal and Triggs 2005).

The rest of the paper is as follows. First we review related work in section 2. Section 3 introduces the actual data pre-processing stage and feature extraction. Machine learning for the purpose of gesture classification is introduced in section 4. Datasets and experimental methodology are explained in section 5. Section 6 presents and discusses the results. Conclusions and future work are drawn in section 7.

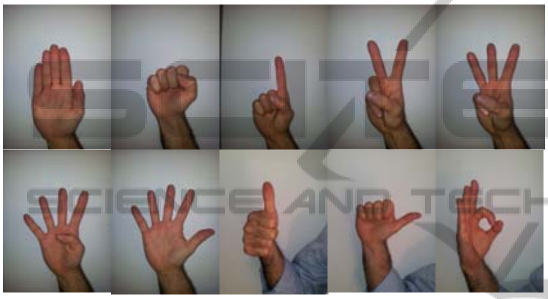


Figure 1: The defined gesture vocabulary.

2 RELATED WORK

Hand gesture recognition is a challenging task in which two main approaches can be distinguished: hand model based and appearance-based methods (Ong and S.Ranganath 2005; Conseil, Bourenname et al. 2007). Although appearance-based methods are view-dependent, they are more efficient in computation time. They aim at recognizing a gesture among a vocabulary, with template gestures learned from training data, whereas hand model-based methods are used to recover the exact 3D hand pose. Appearance-based models extract features that are used to represent the object under study. These methods must have, in the majority of cases, invariance properties to translation, rotation and scale changes. There are many studies on gesture recognition and methodologies well presented in (Mitra and Acharya 2007; Murthy and Jadon 2009). Wang et al. (Wang and Wang 2008) used the discrete Adaboost learning algorithm integrated with SIFT features for accomplishing in-plane rotation invariant, scale invariant and multi-view hand detection. Conceil et al. (Conseil, Bourenname et al. 2007) compared two different shape descriptors, Fourier descriptors and Hu moments, for the recognition of 11 hand postures in a vision based approach. They concluded that Fourier descriptors gives good recognition rates in comparison with Hu moments. Barczak et al. (Barczak, Gilman et al. 2011) performed a performance comparison of Fourier descriptors and geometric moment invariants on an American Sign Language database. The results showed that both descriptors are unable to differentiate some classes in the database. Bourennane et al. (Bourennane and Fossati 2010) presented a shape descriptor comparison for hand posture recognition from video, with the objective of finding a good compromise between accuracy of recognition and computational load for a real-time application. They run experiments on two families of contour-based Fourier descriptors and two sets of region based moments, all of them invariant to translation, rotation and scale-changes of hands. They performed systematic tests on the Triesch benchmark database and on their own with more realistic conditions, as they claim. The overall result of the research showed that the common set Fourier descriptors when combined with the k-nearest neighbour classifier had the highest recognition rate, reaching 100% in the learning set and 88% in the test set. Huynh (Huynh 2009) presents an evaluation of the SIFT (scale invariant feature transform), Colour SIFT, and SURF (speeded up robust features) descriptors on very low resolution images. The performance of the three descriptors are compared against each other on the precision and recall measures using ground truth correct matching data. His experimental results showed that both SIFT and colour SIFT are more robust under changes of viewing angle and viewing distance but SURF is superior under changes of illumination and blurring. In terms of computation time, the SURF descriptors offer themselves as a good alternative to SIFT and CSIFT. Fang et al. (Fang, Cheng et al. 2008) to address the problem of large number of labelled samples, the usually costly time spent on training, conversion or normalization of features into a unified feature space, presented a hand posture recognition approach with what they called a co-training strategy (Blum and Mitchell 1998). The main idea is to train two different classifiers with each other and improve the performance of both classifiers with unlabelled samples. They claim that their method improves the recognition performance with less labelled data in a semi-supervised way. Rayi et al (Tara, Santosa et al. 2012) used the centroid distance Fourier descriptors as hand shape descriptors in sign language recognition. Their test results showed that the Fourier descriptors and the

Manhattan distance-based classifier achieved recognition rates of 95% with small computational latency. Classification involves a learning procedure, for which the number of training images and the number of gestures are important facts. Machine learning algorithms have been applied successfully to many fields of research like, face recognition (Faria, Lau et al. 2009), automatic recognition of a musical gesture by a computer (Gillian 2011), classification of robotic soccer formations (Faria, Reis et al. 2010), classifying human physical activity from on-body accelerometers (Mannini and Sabatini 2010), automatic road-sign detection (Vicen-Bueno, Gil-Pita et al. 2004; Maldonado-Báscon, Lafuente-Arroyo et al. 2007), and static hand gesture classification (Trigueiros, Ribeiro et al. 2012). K-Nearest Neighbour was used in (Faria, Lau et al. 2009; Faria, Reis et al. 2010). This classifier represents each example as a data in d–dimensional space, where d is the number of attributes. Given a test sample, the proximity to the rest of the data points in the training set is computed using a measure of similarity or dissimilarity. In the distance calculation, the standard Euclidean distance is normally used, however other metrics can be used (Witten, Frank et al. 2011). An artificial neural network 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 (Vicen-Bueno, Gil-Pita et al. 2004; Hninn and Maung 2009; Faria, Reis et al. 2010; Stephan and Khudayer 2010) used it in a gesture recognition system, Faria et al (Faria, Reis et al. 2010) used it for the classification of robotic soccer formations, Vicen-Buéno (Vicen-Bueno, Gil-Pita et al. 2004) used it applied to the problem of traffic sign recognition and Stephan et al used it for static hand gesture recognition for human-computer interaction. 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, Lau et al. 2009; Faria, Reis et al. 2010) used it to classify robotic soccer formations and the classification of facial expressions, Ke et al. (Ke, Li et al. 2010) used it in the implementation of a real-time hand gesture recognition system for human robot interaction, Maldonado-Báscon (Maldonado-Báscon, Lafuente-Arroyo et al. 2007) used it for the recognition of road-signs and Masaki et al used it in conjunction with SOM (Self-Organizing Map) for the automatic learning of a gesture recognition mode. Trigueiros et al. (Trigueiros, Ribeiro 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. In this paper all the features extracted are analysed individually with machine learning algorithms to understand their performance and robustness in terms of scale, translation and rotation invariant static hand gesture recognition.

3 PRE-PROCESSING AND FEATURE EXTRACTION

Hand segmentation and feature extraction is a crucial step in computer vision applications for hand gesture recognition. The pre-processing stage prepares the input image and extracts features used later with the classification algorithms.

In the present study, we used seven data sets with different features extracted from the segmented hand. The hand features used for the training datasets are: the radial signature, the radial signature Fourier descriptors, the centroid distance, the centroid distance Fourier descriptors, the histogram of gradients (HoG), the Shi-Tomasi corner detector and the uniform local binary patterns.

For the problem at hand, two types of images obtained with a Kinect camera were used in the feature extraction phase. The first one, the hand grey scale image was used in the HoG operator, the LBP (local binary pattern) operator and the Shi-Tomasi corner detector. The second one, the segmented hand blob, was used in the radial signature and the centroid distance signature after contour extraction.

3.1 Radial Signature

Shape signature is used to represent the shape contour of an object. The shape signature itself is a one-dimensional function that is constructed from the contour coordinates. The radial signature is one of several types of shape signatures.

A simple method to assess the gesture would be to measure the number of pixels from the hand centroid to the edges of the hand along a number of equally spaced radials (Lockton 2002). For the present feature extraction problem, 100 equally spaced radials were used. To count the number of pixels along a given radial we only take into account the ones that are part of the hand, eliminating those that fall inside gaps, like the ones that appear between fingers or between the palm and a finger (Figure 2). All the radial measurements can be scaled so that the longest radial has a constant length. With this measure, we can have a radial length signature that is invariant to hand distance from the camera.

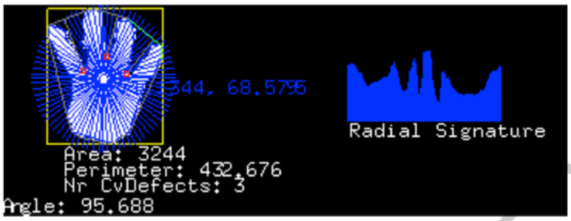


Figure 2: Hand radial signature. Hand with drawn radials (left); obtained radial signature (right).

3.2 Histogram of Gradients (HoG)

Pixel intensities can be sensitive to lighting variations, which lead to classification problems within the same gesture under different light conditions. The use of local orientation measures avoids this kind of problem, and the histogram gives us translation invariance. Orientation histograms summarize how much of each shape is oriented in each possible direction, independent of the position of the hand inside the camera frame (Roth, Tanaka et al. 1998). This statistical technique is most appropriate for close-ups of the hand. In our work, the hand is extracted and separated from the background, which provides a uniform black background, which makes this statistical technique a good method for the identification of different static hand poses, as it can be seen in Figure 3.

This method is insensitive to small changes in the size of the hand, but it is sensitive to changes in hand orientation.

We have calculated the local orientation using image gradients, represented by horizontal and vertical image pixel differences. If $\mathrm{d_x}$ and $\mathrm{d_y}$ are the outputs of the derivative operators, then the gradient direction is $\arctan(\mathrm{d_x}, \mathrm{d_y})$, and the contrast is $\sqrt{d_x^2 + d_y^2}$. A contrast threshold is set as some amount k times the mean image contrast, below which we assume the orientation measurement is inaccurate. A value of k=1.2 was used in the experiments. We then blur the histogram in the angular domain as in (Freeman and Roth 1994), with a [1 4 6 4 1] filter, which gives a gradual fall-off in the distance between orientation histograms.

This feature descriptor was extensively used in many other areas like human detection (Dalal and Triggs 2005; Dalal, Triggs et al. 2006), in conjunction with other operators like the Scale Invariant Feature Transformation (SIFT) (Lowe 2004), the Kanade-Lucas-Tomasi (KLT) feature tracker (Kaaniche and Bremond 2009) and local binary patterns for static hand-gesture recognition (Ding, Pang et al. 2011). Lu et al. (Lu and Little 2006) and Kaniche et al. (Kaaniche and Bremond 2009) used temporal HOGs for action categorization and gesture recognition.

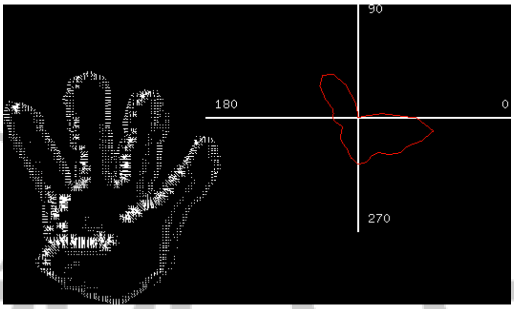


Figure 3: Hand gradients (left), Histogram of gradients (right).

3.3 Centroid Distance Signature

The centroid distance signature is another type of shape signature. The *centroid distance* function is expressed by the distance of the hand contour boundary points, from the centroid (x_c, y_c) of the shape. In our study we used N = 128 as the number of equally sampled points on the contour.

$$\mathrm{d(i)} = \sqrt{[\mathrm{x_i} - \mathrm{x_c}]^2 + [\mathrm{y_i} - \mathrm{y_c}]^2}, \qquad \mathrm{i} = 0, \dots, \mathrm{N} - 1 \tag{1}$$

where $d(i)$, is the calculated distance, and x_i and y_i are the coordinates of contour points. This way, we obtain a one-dimensional function that represents the hand shape.

Due to the subtraction of centroid, which represents the hand position, from boundary coordinates, the centroid distance representation is invariant to translation. Rayi Yanu Tara et al. (Tara, Santosa et al. 2012) demonstrated that this function is translation invariant and that a rotation of that hand results in a circularly shift version of the original image.

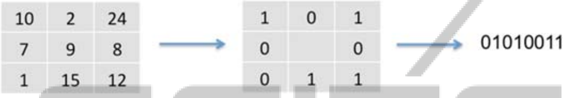
3.4 Local Binary Patterns

LBP (local binary pattern) is a grey scale invariant local texture operator with powerful discrimination and low computational complexity (Ojala, PeitiKainen et al. 2002; Unay, Ekin et al. 2007; Hruz, Trojanova et al. 2011; PietiKainen, Hadid et al. 2011). This operator labels the pixels of the image by thresholding the neighbourhood of each pixel g_0 $(p = 0 \ldots P - 1)$, being P the values of equally spaced pixels on a circle of radius R (R > 0), by the grey value of its center (g_c) and considers the result as a binary code that describes the local texture (Ojala, PeitiKainen et al. 2002; Unay, Ekin et al. 2007; PietiKainen, Hadid et al. 2011). The code is derived as follows:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)\, 2^P \tag{2}$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{3}$$



10	2	24
7	9	8
1	15	12

1	0	1
0		0
0	1	1

01010011

Figure 4: Example of computing $\boldsymbol{LBP_{8,1}}$: example of pixel neighbourhood (left); threshold version (middle); resulting binary code (right).

Figure 4 illustrates the computation of $\mathrm{LBP}_{8,1}$ for a single pixel in a rectangular 3x3 neighbourhood. g_0 is always assigned to be the gray value of neighbor to the right of g_c. In the general definition, LBP is defined in a circular symmetric neighbourhood, which requires interpolation of the intensity values for exact computation. The coordinates of g_0 are given by $(-R\sin(2\pi p/P), R\cos(2\pi p/P))$ (Ojala, PeitiKainen et al. 2002).

The $\mathrm{LBP}_{P,R}$ operator produces 2^P different output values, corresponding to the 2^P different binary patterns that can be formed by the P pixels in the neighborhood set.

As a rotation of a textured input image causes the LBP patterns to translate into a different location and to rotate about their origin, if rotation invariance is needed, it can be achieved by rotation invariance mapping. In this mapping, each LBP binary code is circularly rotated into its minimum value

$$\mathrm{LBP}_{P,R}^{ri} = \min_i \mathrm{ROR}(\mathrm{LBP}_{P,R}\,, i) \tag{4}$$

where $\mathrm{ROR}(x, i)$ denotes the circular bitwise right shift on the P-bit number x, i steps. For example, 8-bit LBP codes 00111100b, 11110000b, and 00001111b all map to the minimum code 00001111b. For P=8 a total of 36 unique different values is achieved. This operator was designated as LBPROT in (Pietikainen, Ojala et al. 2000).

Ojala et.al (Ojala, PeitiKainen et al. 2002) had shown however, that LBPROT as such does not provide very good discrimination. They have observed that certain local binary patterns are fundamental properties of texture, providing the vast majority of all 3x3 patterns presented in observed textures. They called this fundamental patterns “uniform” as they have one thing in common – uniform circular structure that contains very few spatial transitions. They introduced a uniformity measure U(*pattern*), which corresponds to the number of spatial transitions (bitwise 0/1 changes) in the “pattern”. Patterns that have a U value of at most 2 are designated uniform and the following operator for grey-scale and rotation invariant texture description was proposed:

$$\mathrm{LBP}_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_p), & \text{if } U(\mathrm{LBP}_{P,R}) \leq 2 \\ P+1, & \text{otherwise} \end{cases} \tag{5}$$

Equation (5) assigns a unique label corresponding to the number of “1” bits in the uniform pattern, while the non-uniform are grouped under the “miscellaneous” label (P+1). In practice the mapping from $\mathrm{LBP}_{P,R}$ to $\mathrm{LBP}_{P,R}^{riu2}$ is best implemented with a lookup table of 2^P elements. The final texture feature employed in texture analysis is the histogram of the operator output (i.e., pattern labels).

In the present work, we used the histogram of the uniform local binary pattern operator, with R (radius) equal to 1 and P (number of pixels in the neighbourhood) equal to 8, as a feature vector for the hand pose classification.

3.5 Fourier Descriptors

Instead of using the original image representation in the spatial domain, feature values can also be derived after applying a Fourier transformation. The feature vector calculated from a data representation in the transform domain, is called Fourier descriptor (Treiber 2010). The Fourier descriptor is another feature describing the boundary of a region (Snyder and Qi 2004) (Zhang and Lu 2002), and is considered to be more robust with respect to noise and minor boundary modifications. In the present study Fourier descriptors were obtained for the histograms calculated from the radial signature and the centroid distance. For computational efficiency of the FFT, the number of points is chosen to be a power of two (Conseil, Bourenname et al. 2007). The normalized length is generally chosen to be equal to the calculated histogram signature length (N). Hence the Fourier Transform leads to N Fourier coefficients C_k:

$$C_k = \sum_{i=0}^{N-1} z_i \exp\left(\frac{2\pi jik}{N}\right), \qquad k = 0, \dots, N-1 \tag{6}$$

Table 1 shows the relation between motions in the image and transform domains, which can be used in some types of invariance.

Table 1: Equivalence between motions in the image and transform domains.

In the image	In the transform
A change in size	Multiplication by a constant
A rotation of Ø about the origin	Phase shift
A translation	A change in the DC term

The first coefficient C_0 is discarded since it only contains the hand position. Hand rotation affects only the phase information, thus if rotation invariance is necessary, it can be achieved by taking the magnitude of the coefficients. Division of the coefficients by the magnitude of the second coefficient, C_1, on the other hand, achieves scale invariance. This way we obtain N-1 Fourier descriptors I_k:

$$I_k = \frac{|C_k|}{|C_1|}, k = 2, \dots, N-1 \tag{7}$$

Conceil et.al (Conseil, Bourenname et al. 2007), showed that with 20 coefficients the hand shape is well reconstructed, so we used this in our experiments. Centroid distance Fourier descriptors, obtained by applying Fourier transform on a centroid distance signature, were empirically proven to have higher performance than other Fourier descriptors (Zhang and Lu 2002; Shih 2008).

3.6 The Shi-Tomasi Corner Detector

The Shi-Tomasi corner detector algorithm (Shi and Tomasi 1994) is an improved version of the Harris corner detector (Harris and Stephens 1988). The improvement is in how a certain region within the image is scored (and thus treated as a corner or not). Where the Harris corner detector determines the score R with the eigenvalues λ_1 and λ_2 of two regions (the second region is a shifted version of the first one to see if the difference between the two is big enough to say if there is a corner or not) in the following way:

$$R = \det(\lambda_1 \lambda_2) - k(\lambda_1 + \lambda_2)^2 \tag{8}$$

Shi and Tomasi just use the minimum of both eigenvalues

$$R = \min(\lambda_1, \lambda_2) \tag{9}$$

and if R is greater than a certain predefined value, it can be marked as a corner. They demonstrated experimentally in their paper, that this score criteria is much better.

4 MACHINE LEARNING

The study and computer modelling of learning processes in their multiple manifestations constitutes the topic of machine learning (Camastra and Vinciarelli 2008). Machine learning is the task of programming computers to optimize a performance criterion using example data or past experience (Alpaydin 2004). For that, machine learning uses statistic theory in building mathematical models, because the core task is to make inference from sample data.

In machine learning two entities, the teacher and the learner, play a crucial role. The teacher is the entity that has the required knowledge to perform a given task. The learner is the entity that has to learn the knowledge to perform the task. We can distinguish learning strategies by the amount of inference the learner performs on the information provided by the teacher. The learning problem can be stated as follows: given an example set of limited size, find a concise data description (Camastra and Vinciarelli 2008).

In our study, supervised learning was used, where the classification classes are known in advance. In supervised learning, given a sample of input-output pairs, called the training sample, the task is to find a deterministic function or model that maps any input to an output that can predict future observations, minimizing the error as much as possible.

The model was learned from the extracted hand features with the help of the RapidMiner tool. The best learners identified for the produced datasets were the k-NN (k-nearest neighbour), the ANN (artificial neural network) and the SVM (support vector machines).

5 DATASETS & EXPERIMENTAL METHODOLOGY

For data analysis, feature selection, data set preparation and data transformation is an important phase. To construct the right model it is necessary to understand the data. Successful data mining involves far more than selecting a learning algorithm and running it over your data (Witten, Frank et al. 2011). In order to process the recorded videos, a C++ application, using openFrameworks and the respective OpenCV (opencv.org) and OpenNI (openni.org) libraries, was developed. The application runs through all the files, and extracts for each algorithm the respective features. Those features are recorded in text datasets, and converted later to an .xls file to be imported with the Rapid Miner application for data analysis and to find the best learner for each dataset. The experimental results were achieved in an Intel Core i7 (2,8 GHz) Mac OSX computer with 4GB DDR3. All the datasets where analysed with RapidMiner, in order to find the best learner. The experiments were performed under the assumption of the k-fold method. The k-fold cross validation is used to determine how accurately a learning algorithm will be able to predict data that it was not trained with (Camastra and Vinciarelli 2008; Faria, Lau et al. 2009). A value of k=10 (10-fold cross validation) was used, giving a good rule of approximation, although the best value depends on the used algorithm and the dataset (Alpaydin 2004; Witten, Frank et al. 2011).

The algorithms performance, based on the counts of test records correctly and incorrectly predicted by the model, was analysed. Table 2 summarizes the best learners for each dataset with the corresponding parameters.

6 RESULTS AND DISCUSSION

After analysing the different datasets, the obtained results were in most of the cases encouraging, although in other cases weaker than one could expect. In order to analyse how classification errors are distributed among classes, a confusion matrix was computed for each learner with the help of RapidMiner.

Following we present the different results obtained with each dataset, in terms of best learner, the respective confusion matrix and the average accuracy recognition rate.

For the radial signature dataset, the best learner was the neural network with an accuracy of 91,0%. Table 3 shows the obtained confusion matrix. For the centroid distance dataset, the best learner was the neural network, with an accuracy of 90,1%. Table 4 shows the obtained confusion matrix. For the radial signature Fourier descriptors the best learner was the k-NN (nearest neighbour) with a value of k=1 and an accuracy of 82,28%. Table 5 shows the obtained confusion matrix. For the centroid distance Fourier descriptors the best learner was the k-NN (nearest neighbour) with a value of k=1 and an accuracy of 79,53%. Table 6 shows the obtained confusion matrix. For the local binary pattern operator, the best learner was the SVM (support vector machine) with a RBF (radial basis function) kernel type, C = 6 and a bias (offset) of 0.032. The achieved accuracy was 89,3%. The SVM library used was the libSVM (www.csie.ntu.edu.tw/~cjlin/libsvm) (Chang and Lin 2011), since it supports multi-class classification. Table 7 shows the obtained confusion matrix. For the histogram of gradients the best learner was the SVM (support vector machines) with and RBF (radial basis function) kernel type, C = 2 and a bias (offset) of 0.149. The achieved accuracy was of 61,46%. The SVM library used was the libSVM. A lot of misclassification occurred in the data. Table 8 shows the obtained confusion matrix. For the Shi-Tomasi corner detector the best learner was the neural network with a learning rate of 0.1. The obtained results were very weak has can be seen in Table 9.

Table 2: ML algorithms identified as best learners for each dataset and used parameters.

Dataset	Best learn. Algor.	Parameters	Accuracy
Radial Signature	Neural Net		**91,0%**
Centroid Distance	Neural Net		**90,1%**
Radial Sign. Fourier Descriptors	k-NN	k=1	**82,3%**
Centroid dist. Fourier Descriptors	k-NN	k=1	**79.5%**
Uniform Local Binary Patterns	SVM (libSVM)	Kernel = RBF ; C=6; Bias = 0.032	**89,3%**
Histogram of Gradients	SVM (libSVM)	Kernel = RBF ; C=2; Bias = 0.149	**61,46%**
Shi-Tomasi corners	Neural Net	Learning rate = 0.1	**21,90%**

Table 3: Radial signature dataset confusion matrix.

		Actual class									
		1	2	3	4	5	6	7	8	9	10
Predicted class	1	234	1	2	2	3	4	2	6	4	6
	2	2	290	8	2	2	3	1	3	0	6
	3	2	1	273	2	4	5	5	2	2	8
	4	1	1	4	252	6	3	2	4	1	0
	5	5	1	4	2	291	7	1	5	0	0
	6	2	1	2	5	1	281	8	6	2	0
	7	2	1	2	4	1	3	290	3	0	6
	8	2	3	5	3	2	4	0	250	1	5
	9	7	3	9	0	2	3	2	1	276	4
	10	0	8	3	4	4	2	2	2	1	258

Table 4: Centroid distance dataset confusion matrix.

		Actual class									
		1	2	3	4	5	6	7	8	9	10
Predicted class	1	343	7	2	2	5	1	2	2	6	12
	2	9	335	4	4	8	4	12	3	1	1
	3	1	2	314	5	43	0	1	3	0	5
	4	1	0	2	287	7	3	1	12	1	8
	5	2	1	1	2	309	3	8	7	0	9
	6	2	1	0	7	4	345	3	5	9	4
	7	5	4	4	4	0	4	321	1	2	2
	8	3	3	9	3	5	2	1	299	3	3
	9	2	4	6	0	7	3	3	5	308	1
	10	2	4	3	8	11	5	5	9	1	271

Table 5: Radial signature Fourier confusion matrix.

		Actual class									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Predicted class	**1**	250	1	4	2	9	12	3	3	2	2
	2	2	275	10	6	8	3	17	8	1	17
	3	3	5	249	9	7	5	6	17	0	16
	4	7	12	11	248	8	6	7	10	1	0
	5	6	2	4	20	241	16	10	14	2	8
	6	12	3	3	2	21	245	9	4	2	2
	7	3	8	3	5	4	7	228	2	0	10
	8	3	2	13	6	7	9	12	220	1	9
	9	9	1	1	0	2	4	0	6	287	1
	10	1	3	6	0	2	1	6	5	1	232

Table 6: Centroid distance Fourier confusion matrix.

		Actual class									
		1	2	3	4	5	6	7	8	9	10
Predicted class	1	261	17	4	3	13	9	8	5	12	5
	2	8	258	7	8	9	4	8	10	5	6
	3	9	12	295	11	8	2	6	5	6	7
	4	6	6	8	234	7	11	6	7	7	11
	5	2	3	5	6	273	3	12	4	17	17
	6	2	5	4	6	8	290	15	1	6	17
	7	1	6	6	8	3	10	284	12	9	11
	8	9	11	6	5	6	4	3	260	4	14
	9	9	6	7	11	5	7	6	6	242	8
	10	2	6	7	4	7	15	10	15	8	237

Table 7: Local binary patterns dataset confusion matrix.

		Actual class									
		1	2	3	4	5	6	7	8	9	10
Predicted class	1	460	7	4	4	2	2	6	5	10	15
	2	12	499	7	7	8	9	7	3	11	7
	3	2	4	457	24	11	1	2	1	5	9
	4	9	9	12	486	30	6	0	0	8	17
	5	3	15	18	35	522	8	3	0	5	17
	6	10	14	2	4	11	531	4	2	7	15
	7	3	2	1	0	0	1	517	1	2	4
	8	10	1	1	0	0	5	3	554	1	0
	9	5	7	7	8	1	9	4	0	525	4
	10	15	5	31	13	9	4	2	0	1	457

Table 8: Histogram of gradients dataset confusion matrix.

		Actual class									
		1	2	3	4	5	6	7	8	9	10
Predicted class	1	174	15	14	11	5	0	1	17	19	17
	2	24	207	8	11	10	13	8	9	25	12
	3	18	10	199	25	12	4	2	6	20	13
	4	7	5	22	168	24	15	3	4	10	24
	5	8	7	11	18	181	19	15	4	6	19
	6	0	7	2	9	24	195	19	5	7	15
	7	16	39	16	21	34	62	259	39	20	38
	8	10	4	3	5	6	2	1	189	5	3
	9	30	19	17	9	8	3	1	12	176	14
	10	10	15	16	23	13	11	11	3	19	161

Table 9: Shi-Tomasi corner detector confusion matrix.

		Actual class									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Predicted class	**1**	45	26	36	22	23	19	15	17	30	25
	2	22	77	34	16	10	18	16	5	15	59
	3	46	40	49	52	49	40	27	33	25	42
	4	28	30	41	43	48	35	27	23	22	27
	5	23	16	36	36	30	42	22	23	14	11
	6	29	21	39	42	46	56	53	26	29	27
	7	16	23	15	30	34	35	75	16	31	28
	8	27	5	37	23	32	37	16	139	14	9
	9	27	22	12	13	20	26	38	7	104	24
	10	24	58	21	26	23	17	27	7	30	63

7 CONCLUSIONS AND FUTURE WORK

This paper presented a comparative study of seven different algorithms for hand feature extraction, aimed at static hand gesture classification and recognition, for human computer interaction.

We defined our own gesture vocabulary, with 10 gestures (Figure 1), and we have recorded videos from 20 persons performing the gestures for hand feature extraction. The study main goal was to test the robustness of all the algorithms, applied individually to scale, translation and rotation invariance. After analysing the data and the obtained results we conclude that further pre-processing on the video frames is necessary in order to minimize the number of different feature values obtained for the same hand posture. The depth video images obtained with the Kinect have low resolution and some noise, so it was concluded that some imprecision on data recordings results from those problems, leading to more difficult class learning. There are several interpretations of noise as explained in (Alpaydin 2004). Due to this situation, it was decided that a temporal filtering and/or a spatial filtering should be used and will be tested and analysed to see if better results are achieved.

It has been found that the radial signature and the centroid distance are the best shape descriptors discussed in this paper in terms of robustness and computation complexity. Sometimes we have to apply the principle known as *Occam's razor*, which states that "*simpler explanations are more plausible and any unnecessary complexity should be shaved off*".

The Shi-Tomasi corner detector implemented in OpenCV was the one that achieved the weaker results, and we will possibly try it only on future studies with dynamic gestures. Better results were expected from the Fourier descriptors, after having analysed related work on the area, so we will evaluate them further after having implemented the video streaming temporal filtering. In the local binary pattern operator, different radius and number of neighbours will be tested to analyse if better results are obtained.

Also, datasets with a combination of studied features will be constructed and evaluated for the problem at hand.

REFERENCES

Alpaydin, E. (2004). Introduction to Machine Learning, MIT Press.

Barczak, A. L. C., A. Gilman, et al. (2011). Analysis of Feature Invariance and Discrimination for Hand Images: Fourier Descriptors versus Moment Invariants. International Conference Image and Vision Computing. New Zeland.

Ben-Hur, A. and J. Weston (2008). A User's Guide to Support Vector Machines. Data Mining Techniques for the Life Sciences, Humana Press. 609: 223-239.

Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. Proceedings of the eleventh annual conference on Computational learning theory. Madison, Wisconsin, United States, ACM: 92-100.

Bourennane, S. and C. Fossati (2010). "Comparison of shape descriptors for hand posture recognition in video." Signal, Image and Video Processing 6(1): 147-157.

Camastra, F. and A. Vinciarelli (2008). Machine Learning for Audio, Image and Video Analysis, Springer.

Chang, C.-C. and C.-J. Lin (2011). "LIBSVM: A library for support vector machines." ACM Transactions on Intelligent Systems and Technology 2(3): 27.

Conseil, S., S. Bourenname, et al. (2007). Comparison of Fourier Descriptors and Hu Moments for Hand Posture Recognition. 15th European Signal Processing Conference (EUSIPCO). Poznan, Poland: 1960-1964.

Dalal, N. and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection. International Conference on Computer Vision & Pattern Recognition, Grenoble, France.

Dalal, N., B. Triggs, et al. (2006). Human detection using oriented histograms of flow and appearance. 9th European conference on Computer Vision. Graz, Austria, Springer-Verlag 428–441.

Ding, Y., H. Pang, et al. (2011). "Static Hand-Gesture Recognition using HOG and Improved LBP features." International Journal of Digital Content Technology and its Applications 5(11): 236-243.

Fang, Y., J. Cheng, et al. (2008). Hand posture recognition with co-training. 19th International Conference on Pattern Recognition (ICPR 2008). , Tampa, FL.

Faria, B. M., N. Lau, et al. (2009). Classification of Facial Expressions Using Data Mining and machine Learning Algorithms. 4ª Conferência Ibérica de Sistemas e Tecnologias de Informação, Póvoa de Varim, Portugal.

Faria, B. M., L. P. Reis, et al. (2010). Machine Learning Algorithms applied to the Classification of Robotic Soccer Formations ans Opponent Teams. IEEE Conference on Cybernetics and Intelligent Systems (CIS). Singapore: 344 - 349

Freeman, W. T. and M. Roth (1994). Orientation Histograms for Hand Gesture Recognition, Mitsubishi Electric Research Laboratories, Cambridge Research Center.

Gillian, N. E. (2011). Gesture Recognition for Musician Computer Interaction. Doctor of Philosophy, Faculty of Arts, Humanities and Social Sciences.

Harris, C. and M. Stephens (1988). A combined corner and edge detector. The Fourth Alvey Vision Conference.

Hninn, T. and H. Maung (2009). "Real-Time Hand Tracking and Gesture Recognition System Using Neural Networks." 50(Frebuary): 466-470.

Hruz, M., J. Trojanova, et al. (2011). "Local binary pattern based features for sign language recognition." Pattern Recognition and Image Analysis 21(3): 398-401.

Huynh, D. Q. (2009). Evaluation of Three Local Descriptors on Low Resolution Images for Robot Navigation. Image and Vision Computing (IVCNZ '09). Wellington: 113 - 118

Kaaniche, M.-B. and F. Bremond (2009). Tracking HOG Descriptors for Gesture Recognition. IEEE Int. Conf. on Advanced Video and Signal based Surveillance, IEEE Computer Society Press.

Ke, W., W. Li, et al. (2010). "Real-Time Hand Gesture Recognition for Service Robot." 976-979.

Lockton, R. (2002). Hand Gesture Recognition Using Computer Vision, Oxford University.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision 60(2): 91-110.

Lu, W.-L. and J. J. Little (2006). Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor. Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision, IEEE Computer Society: 6.

Maldonado-Báscon, S., S. Lafuente-Arroyo, et al. (2007). Road-Sign detection and Recognition Based on Support Vector Machines. IEEE Transactions on Intelligent Transportation Systems. 8: 264-278.

Mannini, A. and A. M. Sabatini (2010). "Machine learning methods for classifying human physical activity from on-body accelerometers." Sensors 10(2): 1154-1175.

Mitra, S. and T. Acharya (2007). Gesture recognition: A Survey. IEEE Transactions on Systems, Man and Cybernetics, IEEE. 37: 311-324.

Murthy, G. R. S. and R. S. Jadon (2009). "A Review of Vision Based Hand Gestures Recognition." International Journal of Information Technology and Knowledge Management 2(2): 405-410.

Ojala, T., M. PeitiKainen, et al. (2002). "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Trans. Pattern Analysis ans Machine Intelligence 24(7): 971-987.

Ong, S. and S.Ranganath (2005). "Automatic sign language analysis: A survey and the future beyond lexical meaning." IEEE Trans. Pattern Analysis ans Machine Intelligence 27(6): 873-891.

PietiKainen, M., A. Hadid, et al. (2011). Computer Vision Using Local Binary Patterns. London, Springer-Verlag.

Pietikainen, M., T. Ojala, et al. (2000). "Rotation-Invariant Texture Classification using Feature Distributions." Pattern Recognition 33: 43-52.

Roth, M., K. Tanaka, et al. (1998). Computer Vision for Interactive Computer Graphics. IEEE Computer Graphics And Applications, Mitsubishi Electric Research Laboratory: 42-53.

Shi, J. and C. Tomasi (1994). Good Features to Track. Internacional Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, Springer: 593-600.

Shih, F. Y. (2008). Image Processing and Pattern Recognition: Fundamentals and Techniques. Canada, Wiley and Sons.

Snyder, W. E. and H. Qi (2004). Machine Vision, Cambridge University Press.

Stephan, J. J. and S. Khudayer (2010). "Gesture Recognition for Human-Computer Interaction (HCI)." International Journal of Advancements in Computing Technology 2(4): 30-35.

Tara, R. Y., P. I. Santosa, et al. (2012). "Sign Language Recognition in Robot Teleoperation using Centroid Distance Fourier Descriptors." International Journal of Computer Applications 48(2).

Treiber, M. (2010). An Introduction to Object Recognition, Springer.

Trigueiros, P., F. Ribeiro, et al. (2012). A comparison of machine learning algorithms applied to hand gesture recognition. 7ª Conferência Ibérica de Sistemas e Tecnologias de Informação, Madrid, Spain.

Unay, D., A. Ekin, et al. (2007). Robustness of Local Binary Patterns in Brain MR Image Analysis. 29th Annual Conference of the IEEE EMBS, Lyon, France, IEEE.

Vicen-Bueno, R., R. Gil-Pita, et al. (2004). Complexity Reduction in Neural Networks Appplied to Traffic Sign Recognition Tasks.

Wang, C.-C. and K.-C. Wang (2008). Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction. Proceedings of the International Conference on Advanced Robotics (ICAR'07), Jeju, Korea.

Witten, I. H., E. Frank, et al. (2011). Data Mining - Pratical Machine Learning Tools and Techniques, Elsevier.

Zhang, D. and G. Lu (2002). A comparative Study of Fourier Descriptors for Shape Representation and Retrieval. Proc. of 5th Asian Conference on Computer Vision (ACCV), Melbourne, Australia, Springer.