Hand Posture Classification by Means of a New Contour Signature

Nabil Boughnim, Julien Marot, Caroline Fossati, and Salah Bourennane

Institut Fresnel, D. U. de Saint Jérôme Av. Normandie-N.,
13397, Marseille, France
julien.marot@fresnel.fr

Abstract. This paper deals with hand posture recognition. Thanks to an adequate setup, we afford a database of hand photographs. We propose a novel contour signature, obtained by transforming the image content into several signals. The proposed signature is invariant to translation, rotation, and scaling. It can be used for posture classification purposes. We generate this signature out of photographs of hands: experiments show that the proposed signature provides good recognition results, compared to Hu moments and Fourier descriptors.

Keywords: contour description, antenna array, hand posture classification.

1 Introduction

Gesture and posture classification yield many applications in human-computer interaction. We focus on hand posture characterization. Some descriptors have been proposed, which exhibit invariance properties, but skip some details of the hand contour to ensure a low computational load. To overcome this limitation, we seek for a method exhibiting a resolution of one pixel.

1.1 Overview on Hand Posture and Gesture Classification

Systems that employ hand driven human-computer communication interpret hand gestures and postures in different modes of interaction depending on the application domain. Previous works have concentrated on hand gesture classification [1,2], where gesture command is based on slow movements with large amplitude (see for instance in [2] the twelve types of hand gesture). To our knowledge, future applications should concern the classification of hand posture, for the purpose of automated sign language decoding for instance. Contrary to hand gesture, hand posture describes the hand shape and not its movement. Hand posture recognition is a difficult task: the number of 2000 signs is commonly reached in a sign dictionary, so some postures may be very similar.

The main approach for hand posture characterization is based Hu moments [3] and on Fourier descriptors [4]. The advantages of Hu moments and Fourier

J. Blanc-Talon et al. (Eds.): ACIVS 2012, LNCS 7517, pp. 384–394, 2012.

descriptors is their intrinsic invariance to rotation and scaling. For instance, Fourier descriptors were applied for the first time to hand posture recognition in [1]. The illustrations therein show that a large the drawbacks of Hu moments and Fourier descriptors in the context of hand recognition is that a large number of coefficients is required to get an accurate representation of an object. The Fourier descriptors proposed therein provide elevated recognition rates for most of the 11 postures of the considered database. However, Table 4 shows that the algorithm mistakes postures which are visually close. This is due to the low number of used Fourier coefficients -which ensures a low computational load. We wish to distinguish more accurately similar postures, such as the postures 8 and 11. That is why we apply the proposed non star-shaped contour characterization method to distinguish between very similar posture. We predict that the signature provided by the proposed method should yield better classification results than Fourier descriptors, which do not characterize contours with a resolution of one pixel.

1.2 State of the Art on Contour Description and Limitations of Existing Methods

The description of closed contours is a major topic in computer vision. Several features have been proposed: moment invariants in general [5] which aim at extracting shape characteristics independently of scaling, translation and rotation, especially Hu Moment invariants [3], and Fourier descriptors [4]. The main advantage of Fourier descriptors is their invariance to scaling, translation and rotation. Also, their stability has been improved (see [6] and references therein). They involve a regular sampling of the considered contour, the sampling points delimiting arcs of same length. The drawbacks of Fourier descriptors are the following: they are not invariant to the initial description point [6], and a large number of Fourier coefficients, obtained through Fourier transform of the contour coordinates, may be required to distinguish two similar contours (see illustrations in [1]).

Original methods for contour retrieval rely on signal generation on an antenna [7]: A set of virtual sensors forming an antenna, is associated with the image to turn its content into a 1-D signal. These methods were recently improved to detect strongly distorted star-shaped contours [8]. Their main advantages are as follows. First, as opposed to Fourier descriptors, the sampling of the contour does not depend on its shape, but on the chosen directions for signal generation. Refer to [8] for details about signal generation out of the image on a circular antenna. Antenna-based methods permit to distinguish close concentric circles, and their computational load does not depend on the noise level. The main drawback of these antenna-based methods is that they are limited to one-pixel wide star-shaped contours. This prevents the method from fitting various applications like hand posture characterization, considered in this paper.

The limitations cited above lead to the purpose of this paper: we aim at describing a planar free-form contour which may be non star-shaped with a resolution of one pixel.

1.3 Outline of the Paper and Overview of the Proposed Method

For this, we propose a novel scan of the image, inspired from [8] but also from [9]. Various signatures can be associated with contours, for instance distance to the center of mass, complex coordinates, curvature function, or cumulative angles. The image scan in [9] provides a contour signature as a matrix involving the contour polar coordinates. However, it does not offer a pixel-by-pixel description of the contour: its provides the number of pixels in bins which are regularly spaced using some concentric circles and equal interval angle. hence the impossibility to distinguish details which are smaller than the bins. And, the more precise the description, the smaller the regions, but the higher the computational load and the storage place. On the contrary, we propose a contour signature which offers a resolution of one pixel.

The proposed method for image content retrieval splits the image into several rings centered on a reference point. The requirements on the location of this reference point are low, contrary to the condition imposed by the method in [8]. We apply the proposed method to a practical case of non star-shaped contour characterization: hand posture characterization. We aim at distinguishing very similar postures with a computational load which is lower than what the generally used Fourier descriptors would require.

The rest of the paper is organized as follows: in section 3, we define the new representation of contours which is adapted to non star-shaped contours. In section 4, we give a detailed description of the proposed approach for hand posture characterization and report promising results. Concluding remarks and future works are in section 5.

2 Image Acquisition Setup

This setup contains a CMOS camera. Its has the size of a webcam, see Fig. 1 and could further be integrated in an embedded system.

The camera is placed over the desk surface, it axis is orthogonal to the desk surface. Wide angle optics (90°) are used so that the field of vision is wide enough. The acquisition format can be either CIF, or VGA. The video stream is transmitted to the computer by a USB connection in RGB format. The USB connection has a limited stream and requires part of the CPU ressources of the computer.

However at this point, we do not need a high video stream transmission rate. Each video is split into a series of images. The set of images forms our database. Our database contains images with various types of hand postures, which will be described further in detail.

3 A Novel Signature for Non Star-Shaped Contours

A size $N \times N$ image is considered, whose pixels are referred to, starting from the top left corner of the image, as $I_{l,m}$ (see Fig. 2(a)). The 1-valued pixels

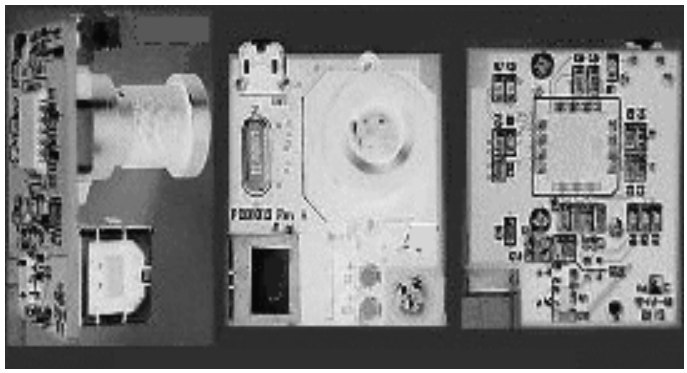


Fig. 1. Camera

compound the expected contour. The contour pixels are located in a system of polar coordinates with pole $\{l_c, m_c\}$ (see Fig. 2(a)). Contrary to the methods proposed in [8], where the center must be chosen in such a way that the contour is star-shaped, the computation of the center coordinates is not essential. What we call signature in this paper is a set of data which characterizes the corresponding contour and permit to reconstruct it. The novel signature that we propose in this paper is based on the generation of signals out of an image. We get inspired from [8] and [9]: as in [8], a circular array of sensors is associated with the image. The sensor array is supposed to be placed along a circle centered on the pole $\{l_c, m_c\}$. The number of sensors is denoted by Q and one sensor corresponds to one direction for signal generation D_i, which makes an angle θ_i with the vertical axis. See for instance the i^{th} and the Q^{th} sensors in Fig. 2(b). The other sensors are not represented for sake of clarity. The method proposed in [8] is valid only for contours exhibiting at most one pixel for one direction D_i. We wish to overcome this limitation and characterize non star-shaped contours. To separate the influence of each pixel located along a given direction D_i, we no longer generate one 1-D signal, but a number L of 1-D signals on the antenna. Each signal corresponds to one 'ring' represented on Fig. 2.

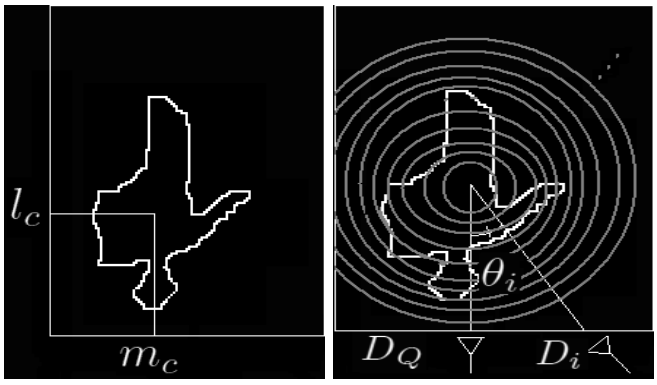


Fig. 2. Image and edge model (a); signal generation process (b)

We assume that, for each direction D_i, there is only one pixel in each of the L intervals. L differs from one direction D_i to another. Its maximum value is, for instance, $\frac{N}{\sqrt{2}}$, if $l_c = N/2$ and $m_c = N/2$.

So, we generate L signal vectors for each direction D_i. For the l^{th} interval ($l = 1, \ldots, L$) and the direction D_i ($i = 1, \ldots, Q$), the signal component $z_{l,i}$ is computed as follows:

$$z_{l,i} = I_{l_{l,i}, m_{l,i}} \sqrt{{l_{l,i}}^2 + {m_{l,i}}^2} \tag{1}$$

The components $z_{l,i}$ can be grouped into a matrix $\mathbf{Z}$ of size $L \times Q$. All columns of $\mathbf{Z}$ should have the same number of rows, so for the directions D_i for signal generation which cross less than L intervals, 0-valued components are set in $\mathbf{Z}$ for the corresponding column indices i. If the width of the intervals is chosen such that there is at most one pixel per direction D_i and per interval, this matrix permits to reconstruct exactly the contour: it contains the radial coordinates of the contour in the system of pole $\{l_c, m_c\}$.

4 Hand Posture Characterization

In this section, we detail the process of hand posture classification: out of a database coming from our acquisition setup, we afford a set of images corresponding to 11 postures (see subsection 4.1). The database is split into a learning database and a test database. But the images cannot be directly exploited: they require a preprocessing which provides the hand contour (see subsection 4.2), so that we can obtain the contour signature as described in section 3. We present the classification process and the two distances that we use for this purpose in subsection 4.3. We provide the results obtained by the proposed method and comparative methods in terms of recognition rates (see subsection 4.4).

4.1 Hand Posture Database

A hand can exhibit a great variety of postures, and it is extremely difficult to recognize all possible configurations of the hand starting from its projection on a 2-D image. Indeed, some parts of the hand can be hidden. It is necessary to consider subsets of postures depending on the application.
There exist some reference databases of specific hand postures, such as the Triesch database [10], available on the Web [11]. This database exhibits limitations: the number of images is low, the viewing angle, the size and the orientation of the hand is always the same, the images are in grey level and contain only the hand. That is why a database made in our research team has been used for experiments in this paper. This database was also used in [12] and in [1]. It contains 11 postures, which are displayed in Fig. 3.

These postures have been chosen to be easily performed by any person. They get inspired from the 8 postures of the completed spoken language presented in [13]. These postures differ from the sign language which aims at easing lip reading. Some postures have been added to test the discrimination performances

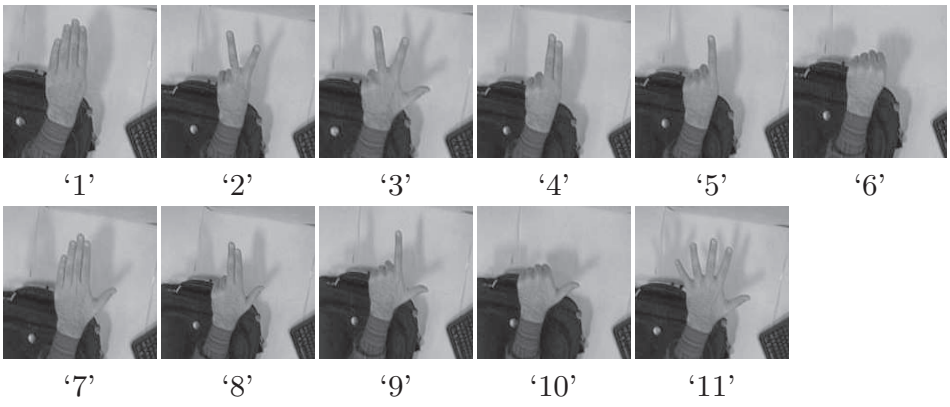


Fig. 3. Postures of the database (cropped images)

of the proposed and comparative methods: they are visually very close, such as postures 4 and 5, as well as postures 8 and 9. The images of the database were obtained as follows: an expert user shoots a movie containing the 11 postures. Then the frames of the video are split to get the images in the RGB color space. These images compound the learning database. Other users which are not the expert user shoot a movie containing the 11 postures. The same process as for the learning database is adopted to get a set of image. These images compound the test set.

In Fig. 4, we display 4 images of the learning database, corresponding to postures 4, 5, 8, and 9.

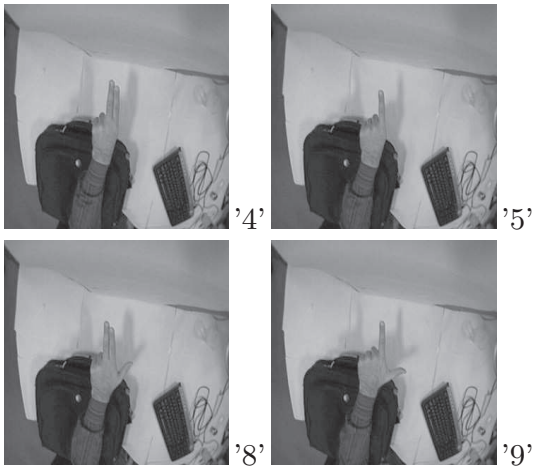


Fig. 4. Images 4, 5, 8, 9 of the database

4.2 Preprocessing

To get images which are fit for hand contour retrieval and posture recognition, we have to get rid of the background and, in the best case, conserve solely the hand contour. In [14], Soriano et al. propose a dynamic skin color model, for a segmentation purpose. Their method copes with changes in illumination. However, their method is applied to faces and not to hands.

We apply the following preprocessing steps to each frame: the color image is mapped to the YC_bC_r space and we select C_b component, where the contrast between hand and background is the largest. We remove the non-moving background from each frame, and an Otsu threshold [15] is applied to the resulting difference frame. Each binary image obtained at this point is impaired by noise; morphological filtering operations of erosion and dilation are applied to eliminate isolated pixels and fill out holes [2].

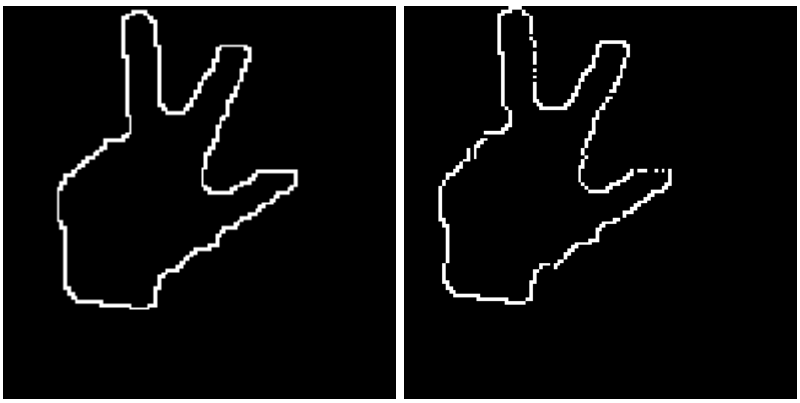
Before getting the image I which is fed to the method proposed in section 3, we apply two further preprocessings.

Firstly, from the initial processed image, we select the smallest subimage containing the expected contour. This subimage is called "enclosing box". The enclosing box is obtained in the following way: the image content is projected onto the left and the bottom sides (it could be also the right and the top sides). We get two signals, $\mathbf{z}^{left}$ and $\mathbf{z}^{bottom}$, from this projection: Their components are obtained as follows: $z_l^{left} = \sum_{m=1}^{N} I_{l,m}\ l = 1, \cdots, N$ and $z_m^{bottom} = \sum_{l=1}^{N} I_{l,m}\ m = 1, \cdots, N$. For each signal, a non-zero section indicates the presence of the expected feature. The l and m indices of the non-zeros sections yield a box enclosing the contour. Extracting this box reduces the computational load of the signature generation.

Secondly, we rotate several times the enclosing box and generate each time $\mathbf{z}^{left}$ and $\mathbf{z}^{bottom}$. We stop when the non zero section length is the largest in $\mathbf{z}^{left}$ and the smallest in $\mathbf{z}^{bottom}$. The hand contour is then straightened up.

Eventually, through the following remarks (•) we can assess that the rows of matrix $\mathbf{Z}$ compose a complete set of invariant features:

• Matrix $\mathbf{Z}$ describes entirely the hand contour: the rows of matrix $\mathbf{Z}$ compose a complete set of invariant features. Fig. 5 illustrates this by showing a segmented hand posture (see Fig. 5(a)), and the contour which is reconstructed out of its signature $\mathbf{Z}$ (see Fig. 5(b)).
• They are invariant to translation: whatever the hand position in the initial image, the box which encloses the contour is blindly estimated.



(a) (b)

Fig. 5. Segmented contour (a); contour reconstructed from the signature $\mathbf{Z}$ (b)

• They are invariant to scaling: whatever the size of the subimage (small number of pixels if the camera is far from the hand, large number of pixels if the camera is near to the hand), the number of intervals for the radial coordinate values L is always the same. Also, the number of directions for signal generation is always the same. This makes the method invariant to scaling. Hence, the signature depends on the shape of the hand, not on its size.
• They are invariant to rotation: whatever the initial orientation of the hand, straightening up the hand contour makes the proposed method invariant to rotation.

Matrix $\mathbf{Z}$ can then be used for classification issues.

4.3 Classification Process

Let's consider H classes of hand postures. For the purpose of hand posture classification, we compare two distances: the Bayesian and the Euclidian distance. For this, we turn any matrix $\mathbf{Z}$ characterizing a posture into a vector $\mathbf{z}$ of size $L * Q$ where $*$ denotes simple multiplication. For each class h, a subset of hand photographs is available. The H subsets compose the learning set. The learning set was created by an expert who knows exactly what position his fingers should have to fit each posture in Fig. 3. A mean invariant vector μ_h and a covariance matrix Λ_h are computed using the learning set for each class. Even if there are small variations from one posture provided by the expert to another, these variations are smoothed through the computation of the mean invariant vector μ_h. Any image coming from the test set and characterized by vector $\mathbf{z}$ is classified by minimizing the Mahalanobis distance:

$$\mathcal{D}_m = (\mathbf{z} - \mu_h)^T \Lambda_h^{-1} (\mathbf{z} - \mu_h) \tag{2}$$

We also provide a study where the distance used for classification is the Euclidian distance: any image coming from the test set and characterized by matrix $\mathbf{Z}$ is classified by minimizing the distance $||\mathbf{Z}_c - \mathbf{Z}||$, where $||.||$ denotes Frobenius norm: $||\mathbf{Z}|| = \sqrt{\sum_{p=1}^{P} \sum_{i=1}^{Q} z_{p,i}^2}$.

The Mahalanobis distance usually provides better classification results, but the Euclidean distance is easier to implement.

4.4 Results and Discussion

We process images of size 320×240 with a 2-core processor @3.2 GHz, using Matlab®. A value $L = 24$ pixels is large enough to get an exclusive signature for each posture and small enough to get a reasonable computational load. To ensure the invariance to scaling, the number of sensors depends only on the maximum size of the enclosing box. The maximum size of the enclosing box is 120×120, so the number of sensors is $Q = \lceil \sqrt{2}\pi 120 \rceil = 534$, large enough to detect the pixels which are the farthest from the image center.

The learning set is composed of 3300 images, and the test set is composed of 440 images.

Table 1 presents the results obtained with either Hu moments, Fourier descriptors, or the proposed method, using Euclidian distance. The results concerning Hu moments and Fourier descriptors are extracted from [1].

Table 1. Recognition rates per posture with internal database and Euclidian distance, on the test set

	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	'10'	'11'	Mean value
HU	79.2	60.3	64.5	60.0	90.8	100	45.3	62.4	62.4	40.6	67.5	66.6
FD1	92.8	75.3	78.4	92.4	93.8	94.5	89.1	70.7	85.5	75.9	76.2	84.1
Proposed method	97.7	89.5	79.1	88.6	75.7	100	86.1	74.4	85.6	87.2	97.0	87.4

Hu moments method exhibits the worst results among the three methods. This methods encounters difficulties with postures 4, 7, and 10. Fourier descriptors exhibit quite good results, but have difficulties for postures 2, 8, and 11.

The recognition rates of Fourier descriptors and the proposed method (see Table 1) confirms the superiority of the proposed method over Fourier descriptors for all postures except postures 4, 5 and 7. So the proposed method outperforms Fourier descriptors for 8 postures out of 11. The mean recognition rate is also higher (87.4 % against 84.1 % for Fourier descriptors). Postures 4, 5 are very similar: if the segmentation is not performed correctly, for instance if the parameters of the mathematical morphology operations are not perfectly tuned, the segmentation of the two joint fingers in posture 4 is very similar to the segmentation of a single finger. Fourier descriptors cannot be used with an acceptable computational time with more than 6 invariant features [12]. So they miss the concave section corresponding to one finger. This is not the case for the proposed method: it yields a pixel-to-pixel precision with a constant computational load, whatever the considered processed image.

Table 2 provides the recognition results obtained while using the Bayesian distance.

Table 2. Recognition rates per posture with internal database and Bayesian distance, on the testing set

	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	'10'	'11'	Mean value
HU	82.3	68.3	66.5	65.0	93.8	89.0	48.3	65.4	82.8	74.4	40.6	70.6
FD1	86.6	90.8	96.4	60.8	97.8	94.3	80.6	64.8	88.6	73.4	96.2	84.6
Proposed method	98.2	92.7	94.1	89.9	84.2	100	88.1	76.4	85.6	97.4	96.9	91.2

Results concerning Hu moments and Fourier descriptors are extracted from [12]. We notice that using the Bayesian distance improves significantly the results obtained by Hu moments. It also slightly improves the results obtained by Fourier descriptors and by the proposed method. Fourier descriptors still get

better recognition results for postures 3, 5, and 9. It has to be privileged for further developments involving the proposed method. While performing our experiments, we noticed that some images from the database yield bad segmentation results, and that these images are particularly concerned with false classification. Sorting and removing these bad segmented images from the learning database could further improve the classification rates. Also, improving the preprocessing step, by changing the mathematical morphology operators for instance, may improve the final recognition results.

Table 3. Proposed and comparative methods, comparison of performances. a) Gabor filtered + PCA + SVM; b) FD1 + Bayesian; c) Proposed method

	'Processing time'	'System performance'	'Interface'	'Recognition rate (%)'
a)	**4 frames/sec**	**3.4 GHz**	**C language**	**93.7**
b)	**20 frames/sec**	**2 GHz**	**C language**	**84.6**
c)	**5 frames/sec**	**3.1 GHz**	**Matlab**	**91.2**

Table 3 shows that we must find a compromise between the recognition rate and the computational load. For our method this compromise is satisfactory. It can even be improved in terms of computational load, if we implement our algorithm in C/C++ language, instead of Matlab. In terms of recognition rates, we could get inspired by the comparative method **a)**. It is based on a preprocessing with Gabor filters. Using Gabor filters to ensure invariance to rotation was proposed in [16] and further developed in [17]. We could also adapt Gabor filtering to improve the preprocessing and thus the recognition rates. Meanwhile, we should pay attention to the compromise between computational load and recognition rate, depending on the final application.

5 Conclusion and Future Works

This paper deals with hand posture classification. A camera acquires photographs of hand postures with a wide viewing angle. We propose a novel signature for the characterization of hand posture. This signature is made of several 1-D signals. Each signal contains radial coordinates of the pixels in an image region which has the shape of a ring. This signature permits to reconstruct the corresponding contour with a precision of one pixel. By applying two preprocessings, we ensure that this signature forms a complete set of features which are invariant to translation, scaling and rotation. This makes this signature fit for hand posture recognition: we associate the proposed signature generation method with Euclidian and Bayesian distances to get recognition results. We reach promising results which outperform the results obtained by Hu moments and Fourier descriptors, for 8 postures out of 11 while comparing with Fourier descriptors. Prospects for this work are as follows: we could use two cameras or a set of cameras, to have access to the parts of the hand which are hidden; also, the preprocessing step could be improved to get a better segmentation result before generating the signature.

Acknowledgements. This work was financially supported by the "Conseil régional Provence Alpes Côte d'Azur", and by the firm Intuisens, to which we are very grateful.

References

1. Bourennane, S., Fossati, C.: Comparison of shape descriptors for hand posture recognition in video. Signal, Image and Video Processing, August 14 (2010) (online First)
2. Zhu, Y., Xu, G., Kriegman, D.J.: A real-time approach to the spotting, representation, and recognition of hand gestures for Human-Computer Interaction. Computer Vision and Image Understanding 85, 189–208 (2002)
3. Hu, M.-K.: Visual pattern recognition by moment invariants. IEEE Trans. on Information Theory 8, 179–187 (1962)
4. Persoon, E., Fu, K.: Shape discrimination using fourier descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence 8(3), 388–397 (1986)
5. Xu, D., Li, H.: Geometric moment invariants. Pattern Recognition 41, 240–249 (2008)
6. Charmi, M.-A., Derrode, S., Ghorbel, F.: Fourier-based geometric shape prior for snakes. Pattern Recognition Letters 29(7), 897–904 (2008)
7. Aghajan, H.K., Kailath, T.: Sensor array processing techniques for super resolution multi-line-fitting and straight edge detection. IEEE Trans. on IP 2(4), 454–465 (1993)
8. Jiang, H., Marot, J., Fossati, C., Bourennane, S.: Strongly concave star-shaped contour characterization by algebra tools. Elsevier Signal Processing (January 2012)
9. Shu, X., Wu, X.-J.: A novel contour descriptor for 2D shape matching and its application to image retrieval. Image and Vision Computing 29, 286–294 (2011)
10. Triesch, J., von der Malsburg, C.: Robust classification of hand postures against complex backgrounds. In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 170–175 (1996)
11. `http://www.idiap.ch/resource/gestures/`
12. Conseil, S., Bourennane, S., Martin, L.: Comparison of Fourier descriptors and Hu moments for hand posture recognition. In: EUSIPCO 2007, Poznan, Poland (2007)
13. Caplier, A., Bonnaud, L., Malassiotis, S., Strintzis, M.: Comparison of 2D and 3D analysis for automated cued speech gesture recognition. In: SPECOM (September 2004)
14. Soriano, M., Martinkauppi, B., Huovinen, S., Laaksonen, M.: Skin detection in video under changing illumination conditions. In: Procs. of the 15th ICPR, vol. 1, pp. 839–842 (2000)
15. Otsu, N.: A threshold selection method from gray level histograms. IEEE Trans. on Systems, Man, and Cybernetics 9, 62–66 (1979)
16. Amin, M.A., Yan, H.: Sign language finger alphabet recognition from Gabor- PCA representation of hand gestures. In: Proceedings of the 6th International Conference on Machine Learning and Cybernetics, Hong Kong, pp. 2218–2223 (2007)
17. Huang, D.-Y., Hu, W.-C., Chang, S.-H.: Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination. Expert Systems with Applications 38, 6031–6042 (2011)