FAST AND IMPROVED HAND CLASSIFICATION USING DIMENSIONALITY REDUCTION AND TEST SET REDUCTION

Nabil BOUGHNIM⋆, Julien MAROT⋆, Caroline FOSSATI⋆, Salah BOURENNANE⋆, Fréderic GUERAULT†

⋆ Groupe GSM, Institut Fresnel, École Centrale Marseille
Domaine Universitaire de Saint Jérôme Av. Escadrille Normandie, 13397, Marseille, France
† Intui-Sense Technologies, Pôle Performance 510 avenue de Jouques F-13400 Aubagne France

ABSTRACT

In this paper, we consider an issue of hand posture classification. We improve a recently proposed signature, a matrix containing the distance of all contour pixels to an arbitrary reference point. Adequate pre-processings ensure the invariance properties of the signature. Candidate postures are preselected with a surface criterion, and Principal Component Analysis (PCA) reduces the dimensionality of the data, which improves the classification process.

***Index Terms*—** Hand posture; hand recognition; classification algorithm; principal component analysis; biometrics.

1. INTRODUCTION

Hand gesture and posture classification is of great interest for human-computer interaction. Previous works have concentrated on hand gesture classification (see references in [1] and [2]), where gesture command is based on slow movements with large amplitude. 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. This task is difficult because each finger must be distinguished and some postures may be similar. The main approach for hand posture characterization is based on moments which are invariant to several image transformations. A review of such moments is available in [3, 4]. Among those moments, Zernike [5] and Legendre [6] moments are based on orthogonal polynomials. For the purpose of hand posture characterization, Hu moments [7] and Fourier descriptors [8] were preferred to other shape descriptors. Hu moments and Fourier descriptors, as other moments, are invariant to translation, rotation and scaling. However, results in [1] show that Fourier descriptors mistake postures which are visually close. This is due to the low number of coefficients, which ensures a low computational load. We wish to improve the rate of recognition, and to reduce the computational load and the memory space.

2. PROBLEM STATEMENT

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 [9], available on the Web [10]. This database exhibits limitations: the number of images is low, the viewing angle, the size and the orientation of the hand are always the same, the images are in grey level and contain only the hand. That is why our database has been used for experiments in this paper. This database was also used in [1]. It contains 11 postures, which are displayed in Fig. 1.

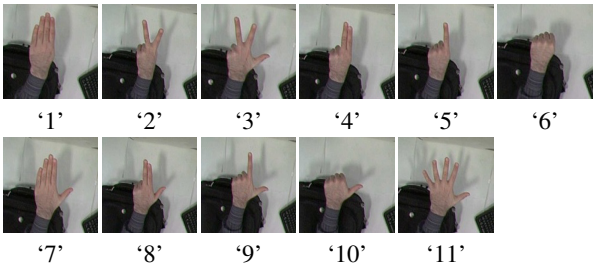

Fig. 1. Postures of the database (cropped images).

These postures have been chosen to be easily performed by any person. These postures differ from the sign language which aims at easing lip reading. Some postures have been added to test the discrimination performances 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 images, these images compound the

test set.
Our goal is as follows: associate any test image with a label, that is, seek for a hand posture recognition method which is invariant to translation, scaling and rotation.

3. DETECTION AND CHARACTERIZATION

For each frame of the learning and test databases, the hand contour must be detected and characterized.

3.1. Detection and preprocessing

The proposed signature generation method requires for each frame a binary image, possibly noise-free. For this, and also to ensure invariance properties, we propose the following preprocessings methods. Firstly, we enhance the contrast between hand and background: the color image is mapped to the YC_bC_r space and we select the C_b component. Secondly, we remove the non-moving background, by substraction of a frame where the hand is not present. Thirdly, we apply an Otsu threshold [11]. Each binary image obtained at this point is impaired by noise; morphological filtering operations -erosions and dilations- are applied to eliminate isolated pixels and fill out holes [2]. These pre-processings also turn the method robust to variations in illumination and inclusion of unexpected objects in the background. We get an image I^f which is supposed to contain only a filled hand.

3.2. Characterization

We assume at this point that we afford an image I, containing the hand contours (see Fig. 2(a)). This image is of size $N \times N$, and its pixels are referred to as $I_{l,m}$, starting from the top left corner. How to obtain I from I^f is explained further. The 1-valued pixel 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)). Each pixel has two coordinates $\{\rho, \theta\}$. The pole can be placed anywhere in the image. In this paper, what we call signature is a set of data which characterizes a contour and permits to reconstruct it. The signature [1] is based on the generation of signals out of an image.

Relation to close prior work

We get inspired from [12] and [13]. In [12] the expected contours are supposed to be star-shaped: whatever the θ value in polar coordinates, there exists only one pixel with coordinate θ. The improvement that we bring in this paper with respect to [12] is a relaxed assumption: the contours no longer have to be star-shaped, which makes the contour characterization method more general. In [13] a histogram of point distribution in a diagram is proposed. This histogram compounds a descriptor. The difference between the proposed signature and the descriptor proposed in [13] is that we reference individually each pixel which is needed to reconstruct the contour.

Firstly, we choose Q directions and a parameter P, the maximum number of intervals for signature generation. Each direction corresponds to an angle θ_q, and each interval is associated with an index p. We assume that, for each direction $D_q,\ q = 1, \ldots, Q$, there is only one pixel in each of the P intervals (see Fig. 2(b)). For some directions D_q the number of intervals is less than P. P is, for instance, $\frac{N}{\sqrt{2}}$, if $l_c = N/2$ and $m_c = N/2$. Secondly, we generate the signature components.

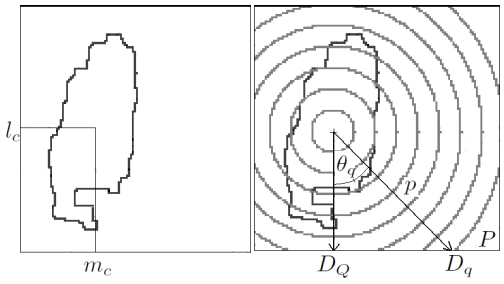

Fig. 2. Image and edge model (a); signal generation process (b).

For the p^{th} interval and the q^{th} direction, a signal component $z_{p,q}$ is computed as follows:

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

The components $z_{p,q}$ can be grouped into a matrix $\mathbf{Z}$ of size $P \times Q$. All columns of $\mathbf{Z}$ should have the same number of rows, so for the directions D_q which cross less than P intervals, 0-valued components are set in $\mathbf{Z}$ for the largest P indices. If the width of the intervals is chosen such that there is at most one pixel per direction D_q and per interval, this matrix permits to reconstruct exactly the $1-$pixel wide contour: it contains the radial coordinates of the contour in the system of pole $\{l_c, m_c\}$. The following two conditions must be fulfilled: $Q \geq \lceil \sqrt{2}\pi N \rceil$, large enough to detect the pixels which are the farthest from the image center; and $\forall\ p$, there exists only one contour pixel in the p^{th} interval. In practice, P is set such that irrelevant pixels (those connected to other pixels, or the isolated noise pixels) are not taken into account.

3.3. Invariance properties

To ensure the invariance to translation of the signature, we select the smallest subimage which contains the hand. This subimage is delimited by an "enclosing box", which is obtained as follows: the content of I^f is projected onto the left and the bottom sides (it could be also the right and the top sides) of the image. This projection is performed as follows: we get two signals, $\mathbf{z}^{left}$ and $\mathbf{z}^{bottom}$ whose components are computed as $z_l^{left} = \sum_m I^f_{l,m}\ l = 1, \cdots, N$ and

$z_m^{bottom} = \sum_l I_{l,m}^f \; m = 1, \cdots, N$. For each signal, a non-zero section indicates the presence of the expected feature. The l and m indices of the non-zero sections yield a box enclosing the contour. The extraction of a subimage ensures the invariance to translation and 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 parameter P and the number of directions for signal generation is always the same. Also, the computational time required for signature generation is reduced.
A linear filter such as prewitt provides the hand contour: we get the contour image I^c. To ensure the invariance to rotation, we rotate several times image I^c in order to consider all possibilities 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 is then straightened up and we afford the image I (see Fig. 2(a)). Matrix $\mathbf{Z}$ forms a complete set of features, which are invariant to translation, scaling, and rotation. It can be used for classification purposes.

4. CLASSIFICATION

Let's consider H classes of hand postures. For the purpose of hand posture classification, Euclidean and Bayesian distances are used in [1]. We will compare the results obtained with Euclidean and Bayesian distances. We vectorize any matrix $\mathbf{Z}$ characterizing a posture into a vector $\mathbf{z}$ of size $P.Q$. For each class h, a subset of hand photographs is available. The H subsets compose the learning set. This set was created by an expert who knows exactly what position his fingers should have to fit each posture in Fig. 1. Let $\mathbf{X}_h$ be the matrix whose columns are the vectors $\mathbf{z}_{n_h},\ n_h = 1, \ldots, M_h$ obtained from the images belonging to class h. For large values of P and Q, $\mathbf{X}_h$ exhibits a large number of rows, and it is a sparse matrix.

4.1. Pre-selection for best posture candidate

To improve the recognition rate, and reduce the computational load and memory space we try to select the most relevant postures of the dictionary. This selection is done through the criterion of hand surface computed from the subimage extracted from I^f. The hand surface is computed for all images of each class in the learning set. Then we choose the following criterion: $|S_t - S_h|$ where S_t is the hand surface for the test image and S_h the mean hand surface for all images of class h in the learning set. Six postures are retained.

4.2. Dimensionality reduction: PCA

Computing the Bayesian distance involves, as shown further in Eq. (2), the inversion of a covariance matrix. To facilitate the computation, we choose to compress the data with principal component analysis (PCA). This permits to reduce the computational load dedicated to matrix inversion, while retaining only the relevant information. Let K be the fixed number of relevant rows in $\mathbf{X}_h$. Let $\mathbf{U}_h$ be the matrix whose columns are the K singular vectors associated with the K largest singular values of $\mathbf{X}_h$. The compressed version of the data is obtained by: $\mathbf{X}_h^c = \mathbf{U}_h^T \mathbf{X}_h$, where ^{T} denotes transpose. Let $\mathbf{z}_{n_h}^c,\ n_h = 1, \ldots, M_h$ denote the columns of $\mathbf{X}_h^c$. The mean invariant vector is computed as $\boldsymbol{\mu}_h = \frac{1}{M_h} \sum_{n_h=1}^{M_h} \mathbf{z}_{n_h}^c$, and the covariance matrix is computed as $\boldsymbol{\Lambda}_h = \frac{1}{M_h} \sum_{n_h=1}^{M_h} \mathbf{z}_{n_h}^c \mathbf{z}_{n_h}^{c\,T}$, for each class $h = 1, \ldots, H$. 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 $\boldsymbol{\mu}_h$. Any image coming from the test set and characterized by vector $\mathbf{z}$ is classified by minimizing the Mahalanobis distance applied to the compressed vector:

$$\mathcal{D}_m = (\mathbf{U}_h^T \mathbf{z} - \boldsymbol{\mu}_h)^T \boldsymbol{\Lambda}_h^{-1} (\mathbf{U}_h^T \mathbf{z} - \boldsymbol{\mu}_h) \qquad (2)$$

for sake of comparaison, the proposed signature can be also exploited with Euclidian distance, computed as follows: $||\mathbf{U}_h^T \mathbf{z} - \boldsymbol{\mu}_h||$, where $||.||$ denotes Frobenius norm.

4.3. Hand posture recognition: summary of the proposed method

Figure 3 presents the overall structure of our algorithm, which improves the recognition rate while requiring a low computational load and a reduced memory space.

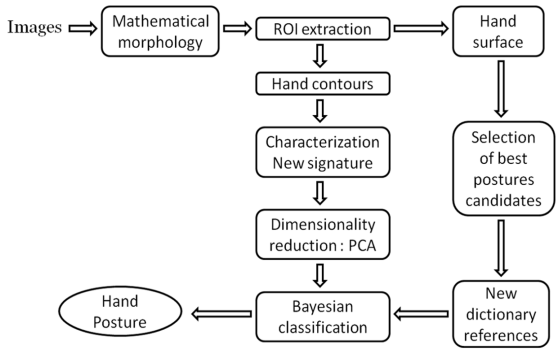

Fig. 3. Improved algorithm for hand posture recognition

5. RESULTS AND DISCUSSION

We process images of size 320×240 with a 2-core processor @3.2 GHz, using Matlab®. A value $P = 24$ levels 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 Q of directions depends only on the maximum size of the enclosing box. This size is 120×120, so the number of directions is set to $Q = \lceil \sqrt{2}\pi 120 \rceil = 534$. The learning set is composed

of 3300 images, and the test set is composed of 495 images. For data compression, the number of relevant features is set to $K = 12$, one more than the number of postures in the dictionary. We present the results obtained with the comparative method Fourier descriptors, and the proposed method, each time using Bayesian distance. We also present some classification rates extracted from [1], where Euclidean distance is used.

	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'	'10'	'11'	OR
A	82.3	68.3	66.5	65.0	93.8	89.0	48.3	65.4	82.8	74.4	40.6	71.1
B	86.6	90.8	96.4	60.8	97.8	94.3	80.6	64.8	88.6	73.4	96.2	84.6
C	97.7	89.5	79.1	88.6	75.7	100.0	86.1	74.4	85.6	87.2	97.0	87.4
D	97.7	100	90.8	86.4	91.7	95.5	93.1	67.4	92.8	97.9	100.0	91.9

Table 1. Comparative results (in %, precision 0.1) obtained with: A) Hu moments with associated with Bayesian distance[7]; B) Fourier Descriptors associated with Bayesian distance[14]; C) the proposed signature associated with Euclidean distance[1]; D) the proposed signature associated with Bayesian distance + PCA. OR is the overall rate of good recognition.

Let's first focus on the overall rate of good classification, indistinctly of the type of posture. It is best for the proposed signature, associated with Bayesian distance + PCA (91.9% versus 71.1%, 87.4% and 84.6%). Then, let's distinguish the postures: we notice that our method is best for postures 2,7,9,10,11. Considering the similar postures 4 and 5, the proposed method provides, though not the best, very good results: 86.4% and 91.7%. Referring to the results obtained with Fourier descriptors and Bayesian distance (see table 1 A), in the easiest cases, the proposed method yields better results than the Fourier descriptors. In the difficult case of posture 4, the recognition rate is much better, and in the difficult case of posture 8, the recognition rate is nearly the same. We consider in the following the proposed signature exploited with Bayesian distance and PCA. Confusion matrix in table 2 shows what are the postures which are mistaken.

	'1'	**'2'**	**'3'**	**'4'**	**'5'**	**'6'**	**'7'**	**'8'**	**'9'**	**'10'**	**'11'**
'1'	**97.7**	0	0	0	0	0	0	0	0	0	0
'2'	0	**100**	0	0	0	0	0	0	0	0	0
'3'	0	0	**90.8**	0	0	0	2.3	4.7	2.4	0	0
'4'	0	0	0	**86.4**	5.5	0	0	0	0	0	0
'5'	2.3	0	2.3	11.3	**91.7**	0	0	0	0	0	0
'6'	0	0	0	0	0	**95.5**	0	0	0	0	0
'7'	0	0	0	0	0	0	**93.1**	2.3	0	0	0
'8'	0	0	0	0	0	0	2.3	**67.4**	2.4	0	0
'9'	0	0	4.5	2.3	2.8	0	2.3	25.6	**92.8**	2.1	0
'10'	0	0	0	0	0	4.5	0	0	2.4	**97.9**	0
'11'	0	0	2.4	0	0	0	0	0	0	0	**100**

Table 2. Confusion matrix (in %, precision 0.1) obtained with: proposed signature, PCA, and Bayesian distance.

The confusion matrix for the proposed classification method (see Table 2) shows that it exhibits good results, except that: posture 4 is recognized as 5 in 11.3 % of the cases, posture 8 is recognized as posture 9 in 25.6 % of the cases; posture 5 as 4 in 5.5 % of the cases. We notice a progress: if Fourier descriptors are used, two postures must be removed from the dictionary to reach a satisfactory overall good classification rate of 90.52 % [14]. If only posture 8 is removed, the proposed method already yields an overall good classification rate of 94.43 %. If both postures 4 and 8 are removed, this rate is up to 95.93 %. We find out that the characterization is not performed correctly for very similar postures as 4,5 and 8,9. This is due to the mathematical morphology operations which yield I^f and follows the detection steps. To avoid mathematical morphology operations, we could improve the detection steps in subsection 3.1 by replacing the YC_bC_r mapping by another process to enhance the hand in the frames.

	'Classif. method'	'Speed'	'System'	'Soft'	'%'	'Database'
a)	PCA+SVM	4 frames/sec	3.4 GHz	C	93.7	11*120
b)	Bayesian	20 frames/sec	2 GHz	C	84.6	11*1000
c)	PCA + Bayesian	6 frames/sec	3.1 GHz	Matlab	91.8	11*45

Table 3. Proposed and comparative methods, comparison of performances. a) Gabor filtered + PCA + SVM [15] ; b) Fourier descriptors (FD1) + Bayesian; c) proposed method.

The computational load dedicated to the recognition of the 495 images of the database is 89.6 sec., that is, a mean rate of 6 frames per second (see Table 3). Fourier descriptors programmed in C++ [14] are faster, namely 25-30 frames per second. However, transferring our programmes from Matlab® to C++ would already decrease the required computational time.

6. CONCLUSION

We improve a hand posture classification method which is based on an original signature. We factor the invariance to translation, scaling, and rotation in the method. In this paper, with the surface pre-selection and dimensionality reduction by PCA included in the proposed method for hand recognition, we improve slightly the recognition rate and computational load. Experiments performed on a large database have shown that the proposed method yields better recognition rates than Fourier descriptors and Hu moments, and that it is faster than a Gabor filter-based method [15]. Our method offers a good compromise between recognition rate and computational load.

Acknowledgements

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

7. REFERENCES

[1] N. Boughnim, J. Marot, C. Fossati, S. Bourennane, "Hand Posture Classification by Means of a New Contour Signature", *ACIVS'12*, vol. 7517, pp. 384-394, 2012.

[2] Y. Zhu and G. Xu, D. J. Kriegman, "A real-time approach to the spotting, representation, and recognition of hand gestures for Human-Computer Interaction", *Computer Vision and Image Understanding*, vol. 85, pp. 189-208, 2002.

[3] M. E. Celebi and Y. A. Aslandogan, "A Comparative Study of Three Moment-Based Shape Descriptors", *ITCC'05*, vol. 1, pp. 788-793, 2005.

[4] A. Foulonneau, P. Charbonnier, F. Heitz, "Geometric shape priors for region-based active contours", *ICIP'03*, vol. 2, pp. 413-416, 2003.

[5] A. Khotanzad, and Y. H. Hong, "Invariant image recognition by Zernike moments", *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 12(5), pp. 489-497, 1990.

[6] M.R. Teague, "Image Analysis Via the General theory of Moments", *J'l of Optical Society of America*, vol. 70(8), pp. 920-930, 1980.

[7] M.-K. Hu, "Visual pattern recognition by moment invariants", *IEEE trans. on Information Theory*, vol. 8, pp. 179-187, 1962.

[8] E. Persoon and K. Fu, "Shape discrimination using fourier descriptors", *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 8, no. 3, pp. 388-397, 1986.

[9] J. Triesch, C. von der Malsburg, "Robust classification of hand postures against complex backgrounds", *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 170-175, 1996.

[10] `http://www.idiap.ch/resource/gestures/`

[11] N. Otsu, "A threshold selection method from gray level histograms," *IEEE trans. on Systems, Man, and Cybernetics*, vol. 9, pp. 62-66, Mar. 1979.

[12] H. Jiang, J. Marot, C. Fossati, S. Bourennane, "Strongly concave star-shaped contour characterization by algebra tools", *Elsevier Signal Processing*, January 2012.

[13] X. Shu, Xi.-J. Wu, "A novel contour descriptor for 2D shape matching and its application to image retrieval", *Image and Vision Computing*, Vol. 29, pp. 286-294, 2011.

[14] S. Conseil, S. Bourennane, L. Martin, Comparison of Fourier Descriptors and Hu Moments for Hand Posture Recognition, 15th European Signal Processing Conference (EUSIPCO'07), Poznan, Pologne, 3-7 Septembre 2007.

[15] D.-Y. Huang, W.-C. Hu, S.-H. Chang, "Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination", *Expert Systems with Applications*, vol. 38, pp. 6031-6042, 2011.