

Mutual Information Based Feature Selection for Fingerprint Identification

Ahlem Adjimi and Abdenour Hacine-Gharbi

LMSE laboratory, University of Bordj Bou Arreridj, Elanasser, 34030 Bordj Bou Arreridj, Algeria

E-mail: adjimia@yahoo.fr, hacgharbi@yahoo.fr

Philippe Ravier

PRISME laboratory, University of Orleans, 12 rue de Blois, 45067 Orléans, France

E-mail: philippe.ravier@univ-orleans.fr, +0033238494863

Messaoud Mostefai

LMSE laboratory, University of Bordj Bou Arreridj, Elanasser, 34030 Bordj Bou Arreridj, Algeria

E-mail: mostefaimess@gmail.com

Keywords: fingerprint identification, feature selection, dimensionality reduction, mutual information, local binary patterns, local phase quantization, histogram of gradients, binarized statistical image features

Received: May 3, 2017

In the field of fingerprint identification, local histograms coding is one of the most popular techniques used for fingerprint representation, due to its simplicity. This technique is based on the concatenation of the local histograms resulting in a high dimension histogram, which causes two problems. First, long computing time and big memory capacities are required with databases growing. Second, the recognition rate may be degraded due to the curse of dimensionality phenomenon. In order to resolve these problems, we propose to reduce the dimensionality of histograms by choosing only the pertinent bins from them using a feature selection approach based on the mutual information computation. For fingerprint features extraction we use four descriptors: Local Binary Patterns (LBP), Histogram of Gradients (HoG), Local Phase Quantization (LPQ) and Binarized Statistical Image Features (BSIF). As mutual information based selection methods, we use four strategies: Maximization of Mutual Information (MIFS), minimum Redundancy and Maximal Relevance (mRMR), Conditional Info max Feature Extraction (CIFE) and Joint Mutual Information (JMI). We compare results in terms of recognition rates and number of selected features for the investigated descriptors and selection strategies. Our results are conducted on the four FVC 2002 datasets which present different image qualities. We show that the combination of mRMR or CIFE feature selection methods with HoG features gives the best results. We also show that the selection of useful fingerprint features can surely improve the recognition rate and reduce the complexity of the system in terms of computation cost. The feature selection algorithms may reach 98% of time reduction by considering only 20% of the total number of features while also improving the recognition rate of about 2% by avoiding the curse of dimensionality phenomena.

Povzetek: Analizirani so različni načini opisa in preiskovanja pri histogramskem kodiranju identifikacije prstnih odtisov.

1 Introduction

Biometric recognition has gained a considerable interest in the recent years because of the various applications in the large field of security. Security can be categorized in data access security (computer and mobile access, USB key, bank cards) or in person access security (forensic identification, ID access). Many technological solutions exist relying on distinctive biometric identifiers (e.g. fingerprints, face, iris or speech) each one having its own qualities. However, the most used biometric identifiers are the fingerprints due to their uniqueness, persistence, simplicity of acquisition and the availability of the electronic acquisition devices [1]. Indeed, the fingerprints are single to each person and they remain unchanged during all the life of the person.

Fingerprint recognition systems can be categorized into three main approaches: minutiae-based systems, image-based correlation systems and image-based distance systems [2]. For the first category, the fingerprint image must pass through several preprocessing steps to detect and extract some points of interest called minutiae: smoothing, local ridge orientation estimation, binarization, thinning, and minutia detection. The second category directly estimates the similarity between a test and a reference fingerprint pattern by the autocorrelation method. For the third category, global or local features are extracted from the fingerprint image such that the features also called descriptors retain most of the pertinent information representing the fingerprint. This kind of

fingerprint recognition systems is preferred in the case of low quality images, because it is difficult to extract reliable minutiae sets in this case [3]. A distance measure between a test and a reference fingerprint pattern or any other classifier are finally used for making a matching decision [3].

Within this last category, many descriptors have been proposed. These descriptors can be principally grouped into histogram-based features or linear transformed features. The descriptors of the first group exploit some statistical characteristics of the fingerprint by transforming the image into a histogram of fixed length like Local Binary Patterns (LBP), Gabor filter with Local Binary Patterns (GLBP) hybrid method [4], Local Phase Quantization (LPQ) [5], Histogram of Gradients [6] or Binarized Statistical Image Features (BSIF) [7] or Scale Invariant Feature Transform (SIFT) [8][9]. In the second group, the fingerprint image is transformed into a vector of different features extracted from the fingerprint image such as Discrete Cosine Transform (DCT) features [10], Gabor filters based descriptors [11][12] and Discrete Wavelet Transform (DWT) features [13][14][15][16].

In this work, we focus on the histogram-based fingerprint representation techniques such as LBP, LPQ, HoG and BSIF. Indeed, these techniques are very used for fingerprint recognition due to their simplicity. These techniques are based on the concatenation of the local histograms leading to a histogram of great dimension (*e.g.* 1024 features for each fingerprint in the case of LBP), which requires long computing time, big memory capacity and requires a huge training dataset to model the classes. Practically, it has been observed that features addition can cause a performance degradation of the classifier if the number of data used for the classifier designing is too low relatively to the number of features [17][18]. This phenomenon called the curse of dimensionality leads to the phenomenon of "peaking" [19]. So it is desirable to keep the number of features as small as possible which is also of benefit for reducing computational cost in the fingerprint identification task and for avoiding memory obstruction too. Keeping a small number of features is a dimensionality reduction operation, which can be done with two approaches: the first approach is a features transformation in which the initial features set is replaced by a new reduced set using transformation algorithm like PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis)... The second approach is a features selection which selects the relevant features from the initial features set [20]. However, using a reduced set of features by transformation needs greater memory capacity and more computing time in the testing phase compared to using a reduced set of features obtained by selection algorithms [20] because the former requires computation of all the features before reduction. So, in the present work, we have considered the features selection algorithms to select the relevant bins of histograms for the histogram-based fingerprint representation techniques.

The feature selection methods are also divided into two categories, which are "wrapper" or "filter". In "wrapper" methods, the relevance measure for a features subset is the

training/testing recognition rate of the used classifier. Consequently, the wrapper selection procedure makes the computational cost rapidly increase, because a new classifier has to be built with training and testing phases each time a features subset is tested. Moreover, the features selected by wrapper methods are adapted to the used classifier, so their performance results are dependent on the type of classifier. In contrast, "filter" methods evaluate the features subset relevance independently of the classifier, so the selected features can be used for any classifier modelling [20][21]. For all these reasons, we have chosen the "filter" methods, which are the preferable methods in the case of high dimensionality and large datasets for computational reasons.

The "filter" methods use a selection criterion typically based on information theory tools like Mutual Information (MI) useful for measuring the quantity of information that features may have for describing the data. To our knowledge, only few works have investigated the MI based criteria in the field of biometric identification.

In [22], an efficient code selection method for face recognition is presented and compact LBP codes are obtained. The code selection is based on the maximization of mutual information (MMI) between features (LBP codes) and class labels. Applying this principle for selection is achieved by using the max-relevance and min-redundancy (mRMR) criterion. The method proposed consists of transforming the face images into LBP histograms, then selecting the relevant codes from these histograms using the maximization of the mutual information. In this work the authors have used the chi-square formula for measuring the distance between the histograms of the reference and the test templates.

In [23], the BSIF features have been investigated in the frame of a fingerprint recognition system, with preliminary results of feature selection using the FVC2002 fingerprint dataset [24]. The experiments have shown that an increasing number of extracted sub-images leads to an increasing recognition rate, but also leads to higher dimension histograms which decreased accordingly performance of the system regarding computing time and memory capacity. This motivated the use of MI feature selection strategy, namely interaction capping (ICAP).

In this work, we extend the fingerprint recognition system proposed in [23] by considering more datasets within the FVC2002 fingerprint database, more descriptor types and by investigating several other feature selection strategies, all based on mutual information computation to select the relevant bins of histograms that are extracted from the fingerprint images. The present study will focus on robustness of the fingerprint system regarding various descriptors and noisy datasets. The main aim of this work is to find a combination of feature selection method with a pertinent descriptor type in a larger context than in study [23]. To that aim, next section introduces the former developments of [23] and explains the novelty of the present paper comparatively. Section 3 proposes a brief review of all the descriptors used in this paper. Section 4 describes the feature selection methods based on mutual information. In section 5 we present the experimental

procedure and we discuss the obtained results using a public fingerprint dataset in section 6. Finally, we draw a conclusion in section 7.

2 Related work

In our previous works [23] and [25], a fingerprint recognition system was created following the flowchart of Fig. 1. A sequence of many preprocessing steps were applied on the training and testing image datasets before extracting the LBP, LPQ or BSIF features, namely enhancement, alignment, extraction of the region of interest (ROI) around the core point and division of the ROI into sub-regions. This procedure is detailed in [23]. So the set of sub-regions are inputs for the features computation. In [25], we used the novel BSIF descriptor [7] compared with LBP and LPQ descriptors, for fingerprint images. From each sub-region, a histogram of BSIF is extracted and the final feature vector is obtained by concatenating all BSIF histograms extracted from the sub-regions. In [23] an extended work of this previous work was presented, in which the relevant bins of the BSIF descriptor extracted histograms were selected using ICAP features selection method. The last step of Fig. 1 is the decision making. It is based on the distance between the histograms of the reference fingerprints and the tested one. The distance is computed as a chi-square measure which formula is given below [22]

$$\chi^2(R, T) = \sum_{i=1}^n \frac{(R_i - T_i)^2}{R_i + T_i} \quad (1)$$

where R_i and T_i are the reference and the tested fingerprint histogram magnitudes respectively and n is the number of bins.

The recognition system uses the following rule to make a decision: if a test fingerprint gives the best match for the fingerprint of the same person it is declared to be a correct match; else it is declared to be a false match.

The recognition rate is computed as

$$\text{Recognition rate (\%)} = \frac{\text{number of correctly recognized images}}{\text{number of test images}} \times 100, \quad (2)$$

In the current paper, many extensions are proposed with respect to our former work [23]. The purpose is to evaluate the robustness of the system regarding changes in the datasets, depending on the descriptors type. We thus consider the new descriptor histogram of gradients (HoG). Then all the descriptors LBP, LPQ, HoG and BSIF are evaluated on all the datasets DB1, DB2, DB3, DB4 of the FVC2002 fingerprint dataset [24]. Indeed, the DB2 and DB3 datasets were discarded for the preliminary study in work [23] while interesting for a robustness study because these are noisy datasets. Moreover, four MI strategies instead of only one in work [23] are investigated for achieving a comparison between them, also by considering the four descriptors instead of BSIF only as proposed in [23]. These novelties are described in the flowchart of Fig. 2. Furthermore, the impact of feature

selection on computing time is analyzed. A deep performance analysis of the dimensionality reduction procedure is also proposed.

The parameter values of the fingerprint recognition system depicted in Fig. 2 will be given in section 5.2 of the experimental part.

3 A brief review of descriptors LBP, LPQ, HoG and BSIF

In this section we give a brief review of the descriptors LBP, LPQ, HoG and BSIF used in this work for features extraction.

3.1 LBP (Local Binary Patterns)

This operator was proposed by Ojala et al [26] for texture analysis. It is characterized by its tolerance to illumination changes, its computational simplicity and its invariance against changes in gray levels. The LBP descriptor works on eight neighbors of a pixel and uses the gray value of this pixel as a threshold; thus, if a neighbor pixel has a higher or a same gray value than the center pixel then a binary one is assigned to that pixel, else it gets a binary zero. The LBP code for the center pixel is then produced by concatenating the eight ones or zeros to obtain a binary number that is transformed after that to a decimal number. The LBP code has a certain value from 0 to 255. Therefore, a histogram of 256 bins is composed from these values and used for matching.

3.2 LPQ (Local Phase Quantization)

This texture descriptor was originally proposed by Ojansivu and Heikkila [27]. It is based on the blur invariance property of the Fourier phase spectrum. It has shown good performance in recognition of textures even when there is no blur and outperforms the Local Binary Pattern operator in texture classification. It uses the local phase information extracted using the 2-D local Fourier transform computed over a window of size $(2R+1)$ by $(2R+1)$ neighborhood at each pixel position in image of size n by n . For LPQ, only four complex coefficients corresponding to 2-D spatial frequencies $v_1 = [a, 0]$, $v_2 = [0, a]$, $v_3 = [a, a]$ and $v_4 = [-a, a]$ where $a = \frac{1}{2R+1}$ are retained. The real and the imaginary parts of the complex values are stacked in a vector of 8 components for each pixel which gives a matrix of size 8 by $n \times n$. Then, the coefficients are decorrelated by a whitening operation assuming a correlation coefficient of 0.95 between adjacent pixel values and a Gaussian distribution of the pixel values. Finally, this matrix is binarized by looking the sign of each element, so that if it has a positive value, a binary 1 is assigned to that element otherwise a binary 0 is assigned. The last step is the histogram construction by transforming each column of 8 elements to a decimal value between 0 and 255. Finally a 256-dimensional histogram is composed from these values and used in classification.

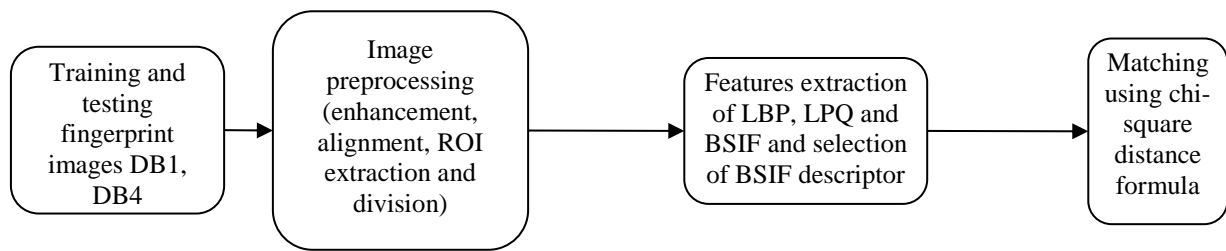


Figure 1: Flowchart of the related work system of fingerprint recognition.

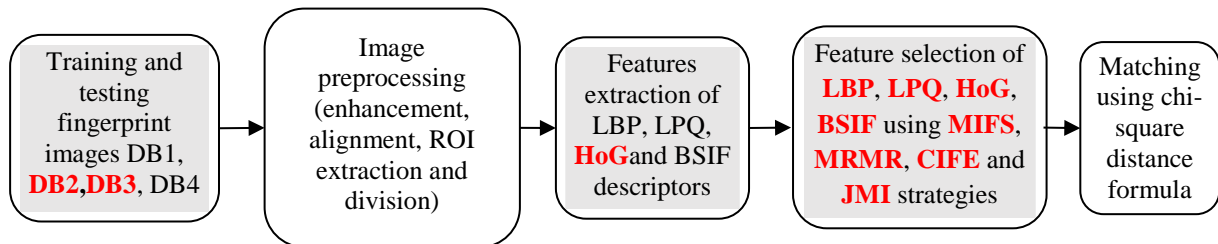


Figure 2: Flowchart of the proposed system. The red characters indicate the added elements for a deep study of the system (details of image preprocessing and matching steps can be found in reference [23]).

3.3 HoG (Histogram of Gradients)

The HoG descriptor has been first proposed by Dalal and Triggs [28] as an image descriptor used in computer vision and image processing for object detection. The basic idea of this descriptor is that local object appearance and shape can be characterized rather well by the distribution of local intensity gradients. The gradient filter is applied in both directions x and y of the image. The two obtained images are then transformed in magnitude and orientation gradients. After, they are divided into small spatial regions (cells). For each cell, each pixel has a gradient magnitude which accumulates the distribution at the bin corresponding to its orientation value. The concatenation of these histograms gives the HoG histogram. For example, if the number of orientation bins spaced over $0^\circ - 180^\circ$ is 9 ($180^\circ/20^\circ$) and the image is split into 3×4 cells (12 is the total number of cells), we then obtain a histogram of G with $3 \times 4 \times 9 = 108$ bins. Actually, the obtained histogram is not a genuine one since the bins cumulative does not reach the total number of pixels. A histogram-like is finally obtained with sqrt L2-normalization [28].

3.4 BSIF (Binarized Statistical Image Features)

BSIF is a new descriptor recently proposed by Kannala & Rahtu [7] for texture classification and face recognition. Its main idea is that it automatically learns a set of filters from a small set of natural images instead of using manual filters such as in LBP and LPQ descriptors. BSIF is a binary code string which length is the number of filters. Each bit of the code string is computed by binarizing the response of the image to a linear filter from the set with a fixed threshold. Given an image patch X of size $l \times l$ pixels and the i linear filter W_i of the same size

from the set of learned filters, the response s_i is obtained by

$$s_i = \sum_{u,v} W_i(u,v)X(u,v) = w_i^T x, \quad (3)$$

where vectors w_i and x contain the pixels of W_i and X . The binarized feature b_i is obtained by setting $b_i = 1$ if $s_i > 0$ and $b_i = 0$ otherwise [7]. The BSIF descriptor depends on two parameters which are the filter window size and the number of bits representing the binary code string. So, the number of bits determines the number of extracted features. If the binary code string is represented with 8 bits, we get 256 features vector, which means a histogram of BSIF features of 256 bins.

4 Feature selection using Mutual Information

Feature selection is used to identify the useful features and remove the features that are redundant and irrelevant for the task of classification. For this reason, it is necessary to reach a measurement of features relevance which makes it possible to quantify their importance in this task. In this section we briefly give some basic concepts and notions from information theory that are useful for understanding the four feature selection methods used in this work. In information theory, MI measures the statistical dependence between two random variables. So, MI can be used to evaluate the relative utility of each feature to classification, in which entropy and mutual information are two principal concepts.

Entropy H can be interpreted as a measure of the uncertainty of random variables. Let X be (or represent) a discrete random variable with probabilistic distribution $p(x)$. The entropy of X is defined as [29]:

$$H(X) = - \sum_{x \in X} p(x) \log(p(x)) \quad (4)$$

The mutual information MI between two discrete variables X and Y is defined using their joint probabilistic distribution $p(x, y)$ and their respective marginal probabilities $p(x)$ and $p(y)$ as:

$$MI(X; Y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (5)$$

The objective of using MI is to select a subset S of relevant features from a set F of features, which share the most information with the class variable. The treatment of each feature needs a very big number of possible subsets (combination C_k^n), this leads to the iterative "greedy" algorithms which select the relevant features one by one (sequential forward selection) or deletes the unneeded features (sequential backward selection). The use of the greedy forward selection procedure with the MI based relevance criterion is generally a good choice of feature selection procedure [30].

The Forward "greedy" algorithm based on MI is presented as follows [31][32]:

- 1) (Initialization) set $F \leftarrow$ "initial set of n features"; $S \leftarrow$ "empty subset"
- 2) (Calculation of MI), $\forall f_i \in F$, calculate $MI(C; f_i)$.
- 3) (Choose the first feature f_{s_1}), find the feature that maximizes $MI(C; f_i)$, affect $F \leftarrow F - \{f_{s_1}\}$, $S \leftarrow \{f_{s_1}\}$.
- 4) (Greedy selection), repeat until the desired number of features:
 - a. (Compute MI between features), $\forall f_i \in F$, compute $MI(C; S, f_i)$.
 - b. (Select the next feature f_{s_j}), choose the feature $f_i \in F$ that maximizes $MI(C; S, f_i)$ at the step j, affect $F \leftarrow F - \{f_{s_j}\}$, $S \leftarrow S \cup \{f_{s_j}\}$.
- 5) Take out the subset S of the selected features.

Practically, it is difficult to compute $MI(C; S, f_i)$ when the cardinal of the subset S increases because it requires an estimation of high dimension probability density functions, which cannot be correctly estimated with a limited number of samples [20]. So the majority of the algorithms use measurements which are maximally based on three variables: two features plus the class index. For this reason, many proposed criteria based on MI are heuristic [32][33].

As previously stated, "filter" methods are preferred to wrapper ones. These methods are defined by a criterion J, also called relevance index or scoring criterion, which is planned to measure the relevance of a feature or a feature subset for the task of classification. The simplest feature-scoring criterion is referred as MIM (Mutual Information Maximization) [21]:

$$J_{mim}(f_i) = MI(C; f_i) \quad (6)$$

The J_{mim} criterion does not include the features already selected which leads to selecting redundant features (sharing the same information with the class index C) that must be eliminated. Numerous "filter" criteria have been proposed taking into account the

redundancy [33][32]. We use four criteria in this work: MIFS, mRMR, CIFE and JMI [21].

4.1 Mutual Information Feature Selection strategy (MIFS)

Proposed by Battiti [31], it is very useful in feature selection problems and classifying systems due to its simplicity. MIFS selects the feature that maximizes the information about the class label C, and subtract the MI between features f_i and the already selected variable f_j to achieve the minimum redundancy:

$$J_{mifs}(f_i) = MI(C; f_i) - \beta \sum_{f_j \in S} MI(f_i; f_j) \quad (7)$$

In this latter expression, S stands for the set of already selected features.

The parameter β is a configurable parameter that determines the degree of redundancy checking within MIFS. It must be set experimentally [21][34]. The performance of MIFS degrades if there are many irrelevant and redundant features because it penalizes redundancy too much.

4.2 Minimum Redundancy and Maximal Relevance strategy (mRMR)

Proposed by Peng et al [35], it is equivalent to MIFS with $\beta = \frac{1}{|S|}$ where $|S| = \text{card}(S)$ is the number of already selected features. It finds a balance between the relevance, which is the dependence between the features and the class, and the redundancy of features with respect to the subset of previously selected features. The criterion can be written as:

$$J_{mrmr}(f_i) = MI(C; f_i) - \frac{1}{|S|} \sum_{f_j \in S} MI(f_i; f_j). \quad (8)$$

With the minimum redundancy criterion of mRMR method, we can get more representative features of the class variable, which are maximally dissimilar to already selected ones, so it gives a small number of features which effectively covers the same space as a larger number of features.

4.3 Conditional Infomax Feature Extraction strategy (CIFE)

Lin and Tang [36] proposed a criterion, called *Conditional Infomax Feature Extraction*, in which the joint class-relevant information is maximized by explicitly reducing the class-relevant redundancies among features [33]. Note that this criterion has been proposed by several authors in different ways [20][32][33][37]:

$$J_{cife}(f_i) = MI(C; f_i) - \sum_{f_j \in S} MI(f_i; f_j) + \sum_{f_j \in S} MI(f_i; f_j | C). \quad (9)$$

	Technology	Scanner	Size of image (pixel × pixel)	Set A	Set B	Resolution
DB1	Optical	IdentixTouchView II	388×374	100 persons with 8 impressions per person (800)	10 persons with 8 impressions per person (80)	500 dpi
DB2	Optical	Biometrika FX2000	296×560			569 dpi
DB3	Capacitive	Precise Biometrics 100 SC	300×300			500 dpi
DB4	Synthetic	SFinGEv2.51	288×384			About 500 dpi

Table 1: The technologies and scanners used to collect the FVC2002 datasets and the size of images in each dataset.

The CIFE criterion is same as MIFS plus the conditional redundancy term.

4.4 Joint Mutual Information strategy (JMI)

Proposed by Yang and Moody [38], the Joint Mutual Information score is

$$J_{\text{jmi}}(f_i) = MI(C; f_i) - \frac{1}{|S|} \sum_{f_j \in S} [MI(f_i; f_j) - MI(f_i; f_j|C)] \quad (10)$$

JMI method studies relevancy and redundancy by taking the mean value, and takes into consideration the class label when calculating MI. JMI and mRMR are very similar but the difference is the conditional redundancy term.

5 Experimental procedure

First, we give a brief description of the public fingerprint dataset FVC2002 [24]. Second, we present the experimental parameters chosen for our fingerprint recognition system. Third, we describe the way we select the relevant bins from LBP, LPQ, HoG and BSIF histograms using the Brown's toolbox for feature selection [21].

5.1 Datasets

The experimental results have been conducted on the FVC2002 fingerprint dataset [24], which has been divided into two sets A and B. Each set is divided in 4 datasets DB1, DB2, DB3 and DB4. Three different scanners and the SFinGe synthetic generator were used to collect the fingerprints [24]. A total of 120 fingers and 12 impressions per finger (1440 impressions) using 30 volunteers have been collected. The top-ten quality fingers were removed from each dataset since they do not constitute an interesting case study [24]. The size of each dataset in the FVC2002 test, however, was established as 110 fingers, 8 impressions per finger (880 impressions) and split into set A (100 fingers - evaluation set) and set B (10 fingers - training set). To make set B representative of the whole dataset, the 110 collected fingers were ordered by quality, and then the 8 images from every tenth finger were included in set B. The remaining fingers constituted set A. In this work, we have used set A to conduct our experimental results [6].

Table 1 presents the technologies and the scanners used to collect the FVC2002 datasets and the size of images in each dataset for each set.

5.2 Fingerprint recognition system

This section describes the experimental parameters chosen for our fingerprint recognition system.

The related work in section 2 mentioned the region around the core point of the fingerprint image. The region of size (100x100 pixels) is extracted and divided into 4 sub-regions of size (50x50 pixels) for each one. For features extraction we use the four descriptors LBP, LPQ, HoG and BSIF applied for each sub-region.

- For LBP features extraction, we convert the gray value of each pixel to one of the 256 LBP codes. Next we construct the histogram of LBP codes.
- For LPQ we use a radius equal to 3, so a histogram of 256 bins is extracted.
- For HoG, each sub-region is divided into sub windows of 3 rows and 3 columns (9 cells total). The orientation and magnitude of each pixel is calculated. The absolute orientation is divided into 9 equally sized bins, which results in a 9-bin histogram per each of the 9 cells, so a histogram of 81 bins is produced.
- For BSIF we use a filter of 11x11 size and number of bits equal to 8 to extract a histogram of 256 bins. The learnt filters are provided by [7].

For each region, the histograms of LBP, LPQ, HoG and BSIF are extracted independently and concatenated to construct the final normalized histogram for each descriptor. The LBP, LPQ, HoG and BSIF histograms are extracted using SifingToolbox¹. For LBP, BSIF and LPQ features, the normalization is carried out by dividing the value of each bin of the histogram by the sum of the values of the bins of this histogram. For HoG features, the normalization is done with sqrt L2-normalization as stated in [28].

Table 2 presents the number of bins in each extracted histogram for the different descriptors.

In this work, the first results are obtained by training the system over 7 images of each person for each dataset. That is, we use 700 dataset images for training and use remaining 100 dataset images for testing for each dataset. In the experiments, the 8 fold-cross validation was applied, so the test step was repeated 8 times.

¹<https://www.dropbox.com/s/wregrs3ah0qcfdd/Sifing.rar>

Feature extraction method	Number of regions around the core point	Number of histogram bins
LBP	4 regions of size 50x50	256*4=1024
LPQ		256*4=1024
HoG		81*4=324
BSIF		256*4=1024

Table 2: Number of histogram bins for each descriptor.

5.3 Bins selection

Table 2 shows that the number of extracted features is high (histogram of 1024 in the case of BSIF, LBP and LPQ and 324 in the case of HoG) which makes the response time in the matching stage very long. The dimensionality reduction is achieved by a feature selection stage. To that aim, we have used the Brown's Toolbox (FEAST toolbox)², which contains the implementation of 13 different features selection methods based on mutual information. In our case we have only used 4 feature selection methods. Two of them are based on the redundancy (MIFS and mRMR). The two other ones are based on the conditional redundancy (CIFE and JMI).

Practically, the LBP, LPQ, BSIF and HoG histogram bins are extracted from all the training images that are also used for feature selection. At this point, each bin is considered as a feature in the feature selection process. This means that each feature is a random variable which probability density function can be estimated with a histogram construction using many realizations of the variable, each image being associated to a realization. Building the histogram of features necessitates the magnitude variation ranges to be properly discretized. This step is required for a low biased estimation of mutual information and entropies used in the Brown's Toolbox. Now, we assume that the number of images is N which is the number of samples or realizations used for histogram estimation of the features. The number m of bins representing the histogram for each feature can be obtained by Sturges' formula [39]:

$$m = \log_2(N) + 1 \quad (11)$$

6 Results and discussion

6.1 Impact of the descriptor type on classification performance

In this section, we analyze performance results of the proposed descriptors for the fingerprint recognition task. Performance is measured in terms of recognition rates and computing time for the identification stage. Table 3 shows the recognition rates and the computing time with all extracted features obtained for each descriptor applied on the different datasets. It is clearly shown from Table 3 (a)

that the LBP features provide the poorest recognition rates compared to the other descriptors in all datasets with an about 10% drop in the recognition rate by comparison with the other rates. The BSIF descriptor gives the best recognition rates except in the DB2 dataset. For all the datasets, the HoG and LPQ descriptors give approximately the same results. It is also observed that DB3 dataset gives the poorest recognition rates. This is due to the fact that DB3 is the most difficult dataset among the four datasets in FVC2002 in terms of image quality [40]. Mainly it can be concluded that the HoG and LPQ descriptors are robust with respect to the dataset diversity because of general high recognition rates compared to the other descriptors. This is confirmed by an average rate over the four datasets reaching near 86.8% for both descriptors. Conversely, BSIF also reaches an average rate of 86% but with extreme values with the highest rates for three datasets and the poorest rate for one dataset. From Table 3 (b), it is clearly shown that the HoG descriptor requires less computing time than the other descriptors for the identification stage. This is due to the smaller number of histogram bins required for this method. Moreover, the computing time is rather independent of the tested dataset. So generally, we can conclude that HoG features outperform the other used features in terms of calculation complexity (only 324 features) and in recognition rate.

A natural perspective is to deal with higher dimension datasets and/or real-time recognition systems. This requires keeping the number of the extracted features as small as possible, which implies computational and memory cost reductions for the training and testing stages. For this reason, many feature selection algorithms have been investigated to solve the problem of computational and memory cost reduction.

6.2 Impact of the feature selection algorithm on classification performance

Fig. 3 shows the results obtained by the four feature selection methods (MIFS, mRMR, CIFE and JMI) on the four datasets DB1, DB2, DB3 and DB4 and with all the descriptors.

The results obtained with LPQ features are very close to those of HoG and BSIF, like observed in the previous study [23] with LBP also giving the poorest results. It can be noted that all the curves reach approximately a plateau as soon as 20% of the total number of features are selected by any of the selection algorithm except MIFS. A first conclusion is that dimensionality feature reduction can be achieved for all the datasets. In many cases, the MIFS algorithm shows an abrupt change at the beginning of the curve. Among the feature selection algorithms, the mRMR is slightly better than the other ones in average over all the datasets.

The curse of dimensionality phenomenon can clearly be observed with DB3 and DB4 datasets in Fig.3, where higher recognition rates can be reached with a smaller number of features than the maximal one. However, the

²<http://www.cs.man.ac.uk/~gbrown/fstoolbox/>

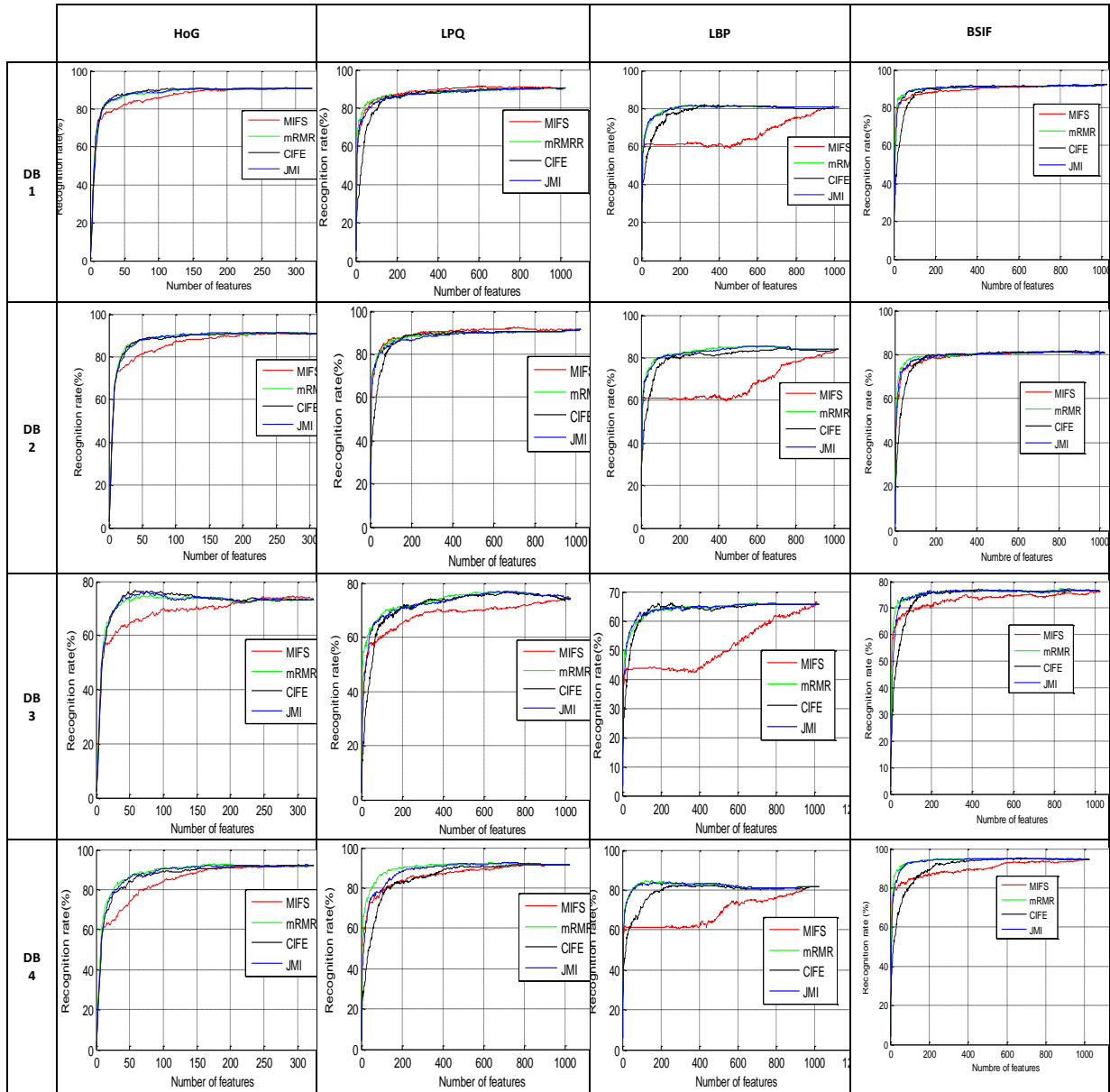


Figure 3: Recognition rates on all the four datasets using HoG, LPQ, LBP and BSIF selected features and using MIFS, mRMR, CIFE and JMI feature selection strategies.

(a)	DB1	DB2	DB3	DB4
HoG	90.75	90.86	73.25	92.13
LPQ	90.25	91.25	74.13	91.50
LBP	80.75	84.00	65.75	81.38
BSIF	92.25	80.75	76.37	94.50

(b)	DB1	DB2	DB3	DB4
HoG	563	569	554	564
LPQ	10350	10493	10304	10378
LBP	10161	10219	10253	10256
BSIF	11381	10905	10609	11257

Table 3: (a) Recognition rate results (%) (b) Computing time results (s) with HoG, LBP, LPQ and BSIF features on the four FVC 2002 datasets.

(a)	DB1	DB2	DB3	DB4
HoG	96.44	96.48	96.39	96.45
LPQ	98.08	98.15	98.11	98.09
LBP	98.08	98.09	98.09	98.11
BSIF	98.01	97.79	97.84	97.94

(b)	DB1	DB2	DB3	DB4
HoG	2.62	2.19	-2.73	4.88
LPQ	4.55	5.2	4.4	3.55
LBP	0	2.98	3.04	-1.99
BSIF	1.76	1.55	0.65	0.66

Table 4: (a) Reduction Rate (%) of computing Time (b) Loss of Recognition Rate (%) caused by dimensionality reduction.

phenomenon of peaking can be far more significant in some curves without cross-validation. Indeed, the curves of Fig.3 are the result of cross-validation which makes an average of 8 recognition rate curves. This operation may mask outlier curves. As an example, we consider a case without cross-validation with HoG features on DB3 by taking the 7th image as a test image and the remainder images as references. From Fig.4, the CIFE algorithm allows 74% of recognition rate to be attained by selecting 28 HoG features which is far better than the recognition rate of 66% obtained with all the features (324). Note in addition that such a case corresponds to the practical use of a feature selection algorithm because of

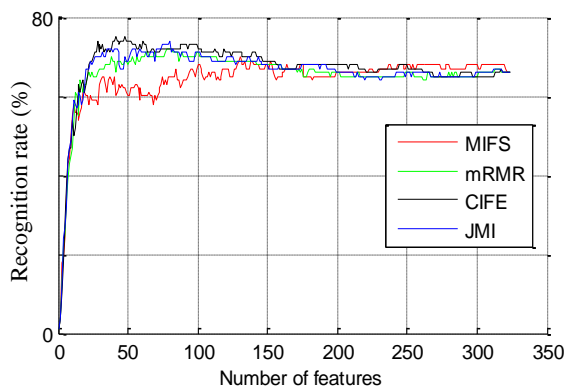


Figure 4: The curse of dimensionality phenomenon (peaking) for DB3 dataset with HoG selected features.

averaging effect of the cross-validation process, which prevents delivering a common sequence of selected features.

6.3 Impact of feature selection on computing time

In this section, we evaluate the benefit of the selection procedure on the complexity of the system in terms of computing time and its effect on the recognition rate of the system. For this experiment, we use the JMI features selection method.

Table 4(a) presents the Reduction Rate of the computing Time (*RRT*) given as follow:

$$RRT = (TF - TS)/TF \tag{12}$$

where *TF* is the computing Time corresponding to number of Full features and *TS* is the computing Time corresponding to the number of Selected features.

Table 4(b) presents the Loss of Recognition Rate (*LRR*) caused by the dimensionality reduction. This is given by:

$$LRR = (RRF - RRS)/RRF \tag{13}$$

where *RRF* is the Recognition Rate corresponding to the number of Full features and *RRS* is the Recognition Rate corresponding to the number of Selected features. In this experiment, we consider the first 20% of the selected features w.r.t. to the full number of features.

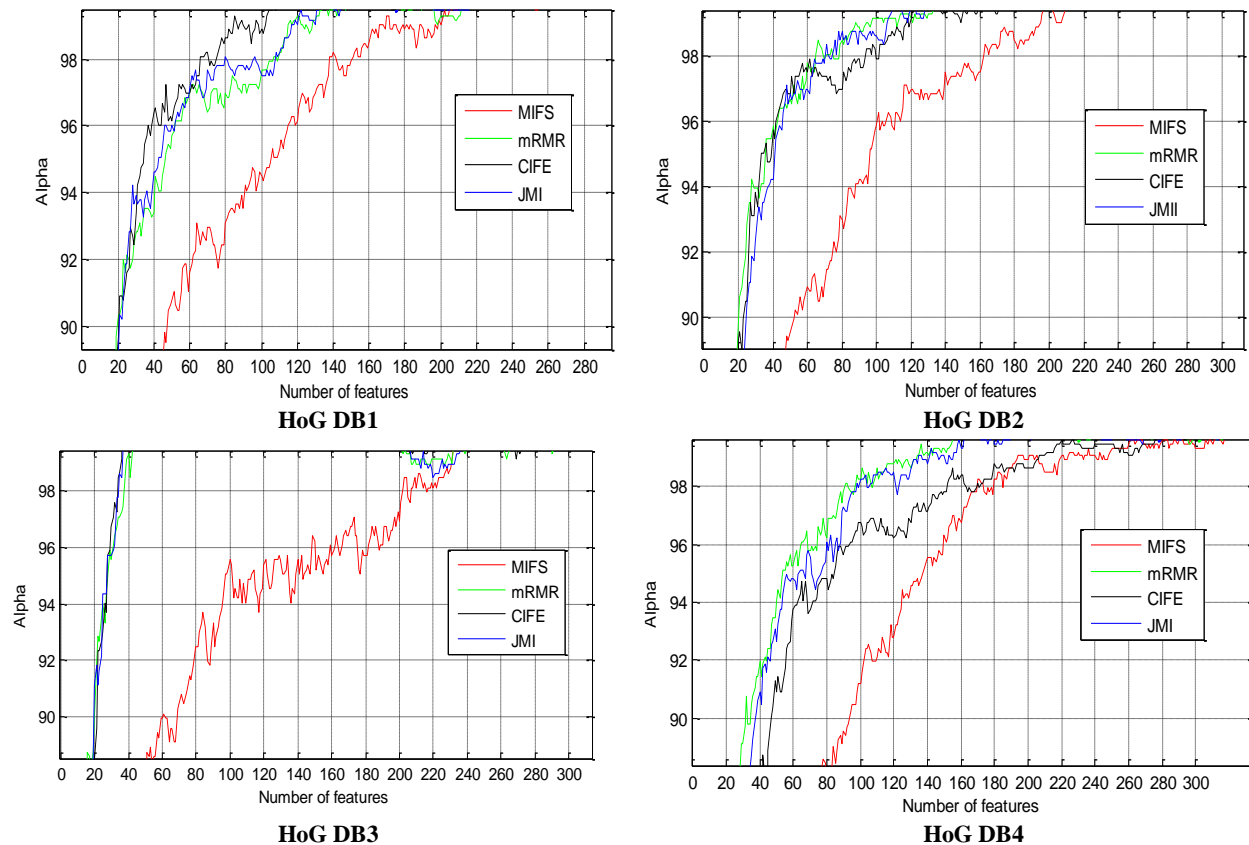


Figure 5: Number of HoG selected features with $\alpha = \{90\% \dots 99\}$ on all datasets, using MIFS, mRMR, CIFE and JMI features selection strategies.

From table 4(a), it can be concluded that considering 20% of BSIF, LBP or LPQ selected features improves the computation time of about 98% compared to the computation time needed with the full number of features. Table 4(b) indicates that the loss of recognition rate may grow up to about 5% while some cases may improve the recognition rate (1.99% when selecting 20% of LBP features with DB3 or 2.73% when selecting 20% of HoG features with DB4 respectively).

6.4 Performance analysis of the dimensionality reduction procedure

It is interesting to know to what extent the number of features could be decreased by considering a small degradation of the recognition rate. For this experiment, we thus determine the number of selected HoG features that allows a recognition rate greater than an α percent value of the rate obtained with the minimum number of features using the formula

$$\alpha = \frac{RRS}{RRF} * 100 \tag{14}$$

where RRS is the recognition rate corresponding to the selected features. RRF is the recognition rate obtained with all the features. The α parameter can take values from 0% to 100%. Fig.5 reports the number of HOG selected features corresponding to α values located in {90%...99%}. From these results, it can be observed that the three feature-selection methods mRMR, CIFE and JMI give very close results, unlike MIFS that always shows poorer performance except in the case of DB3. It can also be observed that CIFE seems to show better results in the case of real bases (DB1, DB2 and DB3) with respect to the synthetic base (DB4). The number of features can be strongly reduced for DB3 with very little concession on the recognition rate (for example 34 features with CIFE are sufficient with $\alpha=98\%$), the profit being very weak for smaller α values. On the other hand, willing to keep the same number of features (34) with the other bases, it is necessary to go down to $\alpha=94\%$ for DB1, 95% for DB2 and less than $\alpha=90\%$ for DB4 (with mRMR).

Table.5 presents the optimal number of BSIF, HoG, LPQ and LBP selected features by the used feature

selection methods with $\alpha=98\%$. Table.6 presents their corresponding recognition rates.

From Tables 5 and 6, the following points can be highlighted:

- For DB1 and DB3, the combination of HoG features with the feature selection method CIFE gives the best performance results with a reduced number of 66 features in the case of DB1 and 34 features in the case of DB3.
- For DB2 and DB4, the combination of HoG features with the feature selection method mRMR gives the best performance results with a reduced number of 66 features in the case of DB2 and 91 in the case of DB4.
- For DB4, using LBP features with feature selection method mRMR gives a reduced number of features equal to 48 but with a poor recognition rate compared to HoG and LPQ. The best performance result is obtained with 87 BSIF features.

As a conclusion, the two feature-selection methods mRMR and CIFE allow obtaining the reduced number of the features in the majority of cases.

7 Conclusion

Histogram based techniques are very used for fingerprint image representation. Generally, concatenation of the histograms leads to the problem of high dimension, which degrades performance results of the identification system in terms of complexity (computing time and memory cost) and recognition rate. In this paper, we have deeply studied the problem of dimensionality reduction in a fingerprint identification system in order to reduce the complexity with possible improvement of the recognition rate avoiding the curse of dimensionality phenomenon. We have presented a fingerprint recognition system based on 4 descriptors: local binary pattern (LBP), local phase quantization (LPQ), Histogram of gradients (HoG) and Binarized Statistical Image Features (BSIF). For the dimensionality reduction we used 4 feature selection methods based on mutual information: MIFS, mRMR, CIFE and JMI. The experiments were conducted on the public FVC 2002 fingerprint dataset.

The use of several types of features and several datasets allows efficiently to validate the feature selection

	BSIF				HoG				LPQ				LBP			
	MIFS	mRMR	CIFE	JMI	MIFS	mRMR	CIFE	JMI	MIFS	mRMR	CIFE	JMI	MIFS	mRMR	CIFE	JMI
DB1	425	202	176	201	138	107	66	80	261	472	313	448	918	144	220	137
DB2	274	113	152	194	162	66	94	75	234	303	255	411	953	207	472	222
DB3	363	121	152	124	202	38	34	35	845	303	290	348	950	260	150	216
DB4	589	90	297	87	170	91	152	98	653	184	425	248	932	48	197	52

Table 5: Number of BSIF, HoG, LPQ and LBP selected features with $\alpha=98\%$. The green values correspond to the minimum number of selected features with a 98% degradation acceptance with respect to the rate obtained with all the features.

	BSIF				HoG				LPQ				LBP			
	MIFS	mRMR	CIFE	JMI	MIFS	mRMR	CIFE	JMI	MIFS	mRMR	CIFE	JMI	MIFS	mRMR	CIFE	JMI
DB1	90	90.37	90.10	90.37	89	89	89	89	88.5	88.5	88.5	88.62	79.38	79.38	79.25	79.38
DB2	79	79.12	79	79.12	89.10	89.5	89.25	89.25	89.5	89.5	89.5	89.5	82.83	82.83	82.5	82.38
DB3	74.74	74.75	74.75	74.75	71.8	72.25	72.10	72.25	72.75	72.75	72.75	72.87	64.5	64.63	64.75	64.5
DB4	92.5	92.5	92.6	92.5	90.5	90.37	90.37	90.30	89.75	89.75	89.87	89.75	79.75	79.75	80.25	79.75

Table 6: Recognition rates obtained by BSIF, HoG, LPQ and LBP selected features with $\alpha=98\%$. The green numbers are those giving the smallest numbers of selected features.

techniques and to choose the best combination (type of features/feature selection method) for the task of fingerprint identification. From all the results we can conclude that the use of feature selection methods can reduce the number of features whatever the type of features and whatever the dataset, except in the case of using MIFS with LBP features that present bad performance result. We can conclude also that the feature selection techniques can reduce the curse of dimensionality phenomenon and probably improve the recognition rate of the identification system. The combination of HoG features with CIFE or mRMR gives the best performance in terms of recognition rate, robustness and complexity of the system. In terms of complexity, a huge computation time reduction (98%) is obtained by considering only 20% of the total number of features without much affecting the recognition rate.

In definitive, employing feature selection algorithms will always provide a benefit when compared to no selection since higher or equal identification performance can be obtained and at the same time the computation complexity for the identification stage can be reduced. As perspective, we plan to investigate other descriptors and biometric modalities.

References

- [1] D. Maio, D. Maltoni, A. K. Jain and S. Prabhakar, "Handbook of fingerprint recognition," Springer, New York, NY, 2003.
<https://doi.org/10.1007/b97303>
- [2] K. S. Sunil, "A Review of Image Based Fingerprint Authentication Algorithms," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, pp. 553-556, 2013.
- [3] Y. Jucheng, "Non-minutiae based fingerprint descriptor," in *Biometrics*, Nanchang, In Tech, 2012, pp. 80-98.
<https://doi.org/10.5772/21642>
- [4] N. Nanni and A. Lumini, "Local Binary Patterns for a hybrid fingerprint matcher," *Pattern Recognition*, vol. 41, no. 11, pp. 3461-3466, 2008.
<https://doi.org/10.1016/j.patcog.2008.05.013>
- [5] S. Brahnam, C. Casanova, L. Nanni and A. Lumini, "A Hybrid Fingerprint Multimatcher," in *16th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 877-882, 2012.
- [6] L. Nanni and A. Lumini, "Descriptors for image-based fingerprint matchers," *Expert Systems With Applications*, vol. 36, no. 10, pp. 12414-12422, 2009.
<https://doi.org/10.1016/j.eswa.2009.04.041>
- [7] J. Kanala and E. Rahtu, "BSIF: binarized statistical image features", *21st International Conference on Pattern Recognition (ICPR 2012)*, IEEE, Tsukuba, Japan, pp. 1363-1366, 2012.
- [8] A I Awad and K Baba, "Evaluation of a fingerprint identification algorithm with SIFT features," in *IIAI International Conference on Advanced Applied Informatics*, Fukuoka, 2012, pp. 129-132.
<https://doi.org/10.1109/iaai-aaai.2012.34>
- [9] S Egawa, A I Awad, and K Baba, "Evaluation of acceleration algorithm for biometric identification," *Network Digital Technologies NDT 2012. Communications in Computer and Information Science. Springer, Berlin, Heidelberg*, vol. 294, pp. 231-242, 2012.
https://doi.org/10.1007/978-3-642-30567-2_19
- [10] T. Amornraksa and S. Tachaphetpiboon, "Fingerprint recognition using DCT features," *Electronic Letters*, vol. 42, no. 9, pp. 522–523, 2006.
<https://doi.org/10.1049/el:20064330>
- [11] A. K. Jain, S. Prabhakar, L. Hong and S. Pankanti, "Filterbank-based fingerprint matching," *Image Processing, IEEE Transactions*, vol. 9, no. 5, pp. 846-859, 2000.
<https://doi.org/10.1109/83.841531>
- [12] S. Lifeng, Z. Feng and T. Xiaoou, "Improved fingenr code for filterbank-based fingerprint matching," in *International Conference on Image Processing*, vol. 2, no. 2, pp. 895-898, 2003.
<https://doi.org/10.1109/icip.2003.1246825>
- [13] R. Kumar, P. Chandra and M. Hanmandlu, "Fingerprint Matching Based on Texture Feature," *In Mobile Communication and Power Engineering*, Springer-Verlag Berlin, vol. 296, pp. 86-91, 2013.
https://doi.org/10.1007/978-3-642-35864-7_12
- [14] M. Saha, J. Chaki and R. Parekh, "Fingerprint Recognition using Texture Features," *International Journal of Science and Research*, vol. 2, no. 12, pp. 2319-7064, 2013.
- [15] K. Tewari and R. L. Kalakoti, "Fingerprint Recognition Using Transform Domain Techniques," in *International Technological Conference*, pp.136-140, 2014.
- [16] M. W. Zin and M. M. Sein, "Texture feature based fingerprint recognition for low quality images Texture Feature based Fingerprint Recognition for Low Quality Images," in *Micro-NanoMechatronics and Human Science (MHS), International Symposium*, IEEE, Nagoya, Japan, pp. 333–338, 2011.
<https://doi.org/10.1109/mhs.2011.6102204>
- [17] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.
- [18] A. K. Jain and B. Chandrasekaran, "Dimensionality and Sample Size Considerations in Pattern Recognition Practice," in *Handbook of Statistics*, Amsterdam, 1982, pp. 835-855.
[https://doi.org/10.1016/s0169-7161\(82\)02042-2](https://doi.org/10.1016/s0169-7161(82)02042-2)
- [19] A. K. Jain, R. P. Duin and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, 2000.
<https://doi.org/10.1109/34.824819>
- [20] A. Hacine-Gharbi, M. Deriche, P. Ravier and T. Mohamadi, "A new histogram-based estimation

- technique of entropy and mutual information using mean squared error minimization," *Computers & Electrical Engineering*, vol. 39, no. 3, pp. 918–933, 2013.
<https://doi.org/10.1016/j.compeleceng.2013.02.010>
- [21] G. Brown, A. Pocock, M. Lujan and M. J. Zhao, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.
- [22] B. Jun, T. Kim and D. Kim, "A compact local binary pattern using maximization of mutual information for face analysis Pattern Recognition," *Pattern Recognition*, vol. 44, pp. 532–543, 2011.
<https://doi.org/10.1016/j.patcog.2010.10.008>
- [23] A. Adjimi, A. Hacine-Gharbi, P. Ravier and M. Mostefai, "Extraction and selection of binarised statistical image features for fingerprint recognition," *Int. J. Biometrics*, vol. 9, no. 1, p. 67–80., 2017.
<https://doi.org/10.1504/ijbm.2017.10005054>
- [24] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman and A. K. Jain, "FVC2002: Second Fingerprint Verification Competition," in *16 th international conference in Pattern Recognition*, 2002.
- [25] A. Adjimi, A. Hacine-Gharbi and M. Mostefai, "Application of Binarized Statistical Image Features for Fingerprint Recognition," in *SIVA 2015, 3rd international conference signal image vision and their applications*, Guelma, Algeria, 2015.
- [26] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 24, no. 7, pp. 971–987, 2002.
<https://doi.org/10.1109/tpami.2002.1017623>
- [27] T. Ojala, M. Pietikainen and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.
[https://doi.org/10.1016/0031-3203\(95\)00067-4](https://doi.org/10.1016/0031-3203(95)00067-4)
- [28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, USA, p. 886–893, 2005.
<https://doi.org/10.1109/CVPR.2005.177>
- [29] T. Cover and J. Thomas, *Elements of information theory*, 2e édition ed., Canada: John Wiley & Sons, 2006.
<https://doi.org/10.1002/047174882x>
- [30] D. François, F. Rossi, V. Wertz and M. Verleysen, "Resampling methods for parameter-free and robust feature selection with mutual information," *Neurocomputing*, vol. 70, pp. 1276–1288, 2007.
<https://doi.org/10.1016/j.neucom.2006.11.019>
- [31] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.
<https://doi.org/10.1109/72.298224>
- [32] A. Hacine-Gharbi, P. Ravier and T. Mohamadi, "Une nouvelle méthode de sélection des paramètres pertinents : application en reconnaissance de la parole," in *conférence TAIMA*, Hammamet, Tunisie, pp. 399–407, 2009.
- [33] G. Brown, "A new perspective for information theoretic feature selection," in *International Conference on Artificial Intelligence and Statistics*, Florida, USA, pp. 49–56, 2009.
- [34] N. Kwak and C. H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.
<https://doi.org/10.1109/72.977291>
- [35] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: Criteria of max dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
<https://doi.org/10.1109/tpami.2005.159>
- [36] D. Lin and X. Tang, "Conditional infomax learning: An integrated framework for feature extraction and fusion," in *European Conference on Computer Vision*, Springer-Verlag Berlin, Graz, Austria, pp. 68–82, 2006.
https://doi.org/10.1007/11744023_6
- [37] I. Kojadinovic, "Relevance measures for subset variable selection in regression problems based on k-additive mutual information," *Comput. Statist. Data Anal.*, vol. 49, pp. 1205–1227, 2005.
<https://doi.org/10.1016/j.csda.2004.07.026>
- [38] H. Yang and J. Moody, "Data Visualization and Feature Selection: New Algorithms for Non Gaussian Data," *Advances in Neural Information Processing Systems*, MIT Press, pp. 688–695, 1999.
- [39] H. Sturges, "The choice of a class-interval," *J. Amer. Statist. Assoc.*, vol. 21, pp. 65–66, 1926.
<https://doi.org/10.1080/01621459.1926.10502161>
- [40] Y. Chen, S. C. Dass and A. K. Jain, "Fingerprint Quality Indices for Predicting Authentication Performance," in *Audio- and Video-Based Biometric Person Authentication*, Springer-Verlag Berlin Heidelberg, Hilton Rye Town, USA, pp. 160–170, 2005.
https://doi.org/10.1007/11527923_17