New COLD Feature based Handwriting Analysis for Enthnicity/Nationality Identification

[1]Sauradip Nag, [2]Palaiahnakote Shivakumara, [3]Yirui Wu, [4]Umapada Pal, and [5]Tong Lu

[1]Kalyani Government Engineering College, Kalyani, Kolkata, India

[2]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[3]College of Computer and Information, Hohai University, Nanjing, China

[4]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India

[5]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

Email: sauradipnag95@gmail.com, shiva@um.edu.my, wuyirui@hhu.edu.cn, umapada@isical.ac.in, lutong@nju.edu.cn,

Abstract—**Identifying crime for forensic investigating teams when crimes involve people of different nationals is challenging. This paper proposes a new method for ethnicity (nationality) identification based on Cloud of Line Distribution (COLD) features of handwriting components. The proposed method, at first, uses tangent angle of the contour pixels in each row and the mean of intensity values of each row for segmenting text lines. For segmented text lines, we use tangent angle and direction of base lines to remove rule lines in the image. We use polygonal approximation for finding dominant points for contours of edge components. Then the proposed method connects the nearest dominant points of every dominant point, which results in line segments of dominant point pairs. For each line segment, the proposed method estimates angle and length, which gives a point in polar domain. For all the line segments, the proposed method generates dense points in polar domain, which results in COLD distribution. As character component shapes change, according to nationals, the shape of the distribution changes. This observation is extracted based on distance from pixels of distribution to Principal Axis of the distribution. Then the features are subjected to an SVM classifier for identifying nationals. Experiments are conducted on a complex dataset, which show the proposed method is effective and outperforms the existing method.**

Keywords—Soft biometric, Nationals identification, COLD features, Handwriting analysis, Ethnicity identification.

I. INTRODUCTION

Traits, namely, gender, nationality, age, height, gait, etc., are popular in the field of biometric applications, such as face and iris recognition [1, 2]. This is because traits prediction helps biometric methods to improve their performances by reducing the complexity of the problem [3, 4]. In addition, traits prediction plays a vital role for forensic applications and security by helping in identifying suspicious behaviors [1]. However, it is noted from biometric based methods that when the input images are posed to an open environment, the methods lose accuracy. This is due to inherent limitations of biometric based methods, such as sensitivity to non-uniform illumination effect, occlusion, degradation, and environment influences. Besides, the methods are said to be computationally expensive as it involves high-level image processing tasks. As a result, in order to help forensic applications and investigation teams, handwriting analysis has become popular, which has now reached beyond traditional boundaries such as emotions, nationality, gender, age and other traits prediction [4, 5]. However, due to large variations in handwriting, ink, pen, paper,

script, age, gender, and individual difference, it is not so easy to identify traits based on handwriting analysis accurately. Therefore, this work focuses on nationality/ethnicity identification as it is useful for identifying crimes where different nationals are involved.

Despite the problem is challenging as mentioned above, one can expect differences in writing styles of different nationals. For example, Chinese usually prefer to write letters with more straight than cursive. This is valid because of the nature of their national language, where alphabets and text usually are formed with the combination of straight strokes. In case of people originating from India and Bangladesh, we can expect more cursive writings than straightness compared to Chinese because most of Indian and Bangladesh scripts are cursive in nature. With this notion, one can confirm that English writing changes from one national to another. Note that the proposed work considers the persons who born in the respective countries but not migrants and non-immigrants. This is the main basis for proposing the method in this work. It is evident from the sample images of each nation shown in Fig. 1, where it can be seen that each nation has a different writing style. At the same time, since all the citizens of respective nations follow their own scripts, we can expect English writing by different persons of the same nation share the common properties.
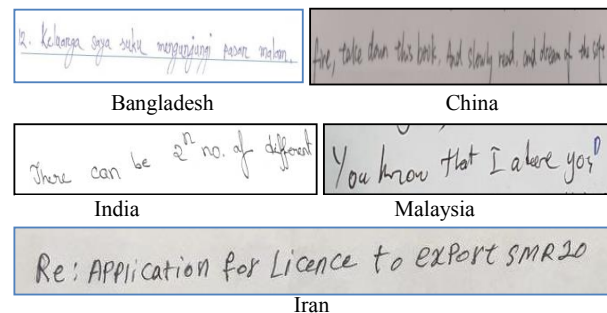

Fig. 1. Examples of English handwriting of different countries

Hardly, we find a method that addresses the issue of nationality and ethnicity identification using handwriting analysis. Maadeed and Hassaine [6] proposed automatic prediction of age, gender, and nationality in offline handwriting. The method extracts direction, curvature, tortuosity, chain code and edge based directional features for character components. They believe that handwriting documents written by different

nationals exhibit the unique difference. Based on this observation, the method extracts the above-mentioned features, which are then passed to a Random Forest classifier for the final classification. This work considers eight nationals for identification. However, the method reports poor results for nationality identification because the proposed features are inadequate to handle intra class variations.

It is noted from the above review that the problem of nationality/ethnicity identification is at infancy stage and hence there is a huge gap from 2014 to 2018. This fact motivated us to propose a new method for nationality identification in this work. The scope of this work is limited to five nationals, namely, Bangladesh, China, India, Malaysia and Iran. The reason to choose these nationals is that we felt English documents written by these nationals are challenging because they share the common ASIAN culture .In addition, we can see most of the above-mentioned countries have multi-religion, multi-lingual, etc. Therefore, this paper presents a method based on the fact that when nations have different scripts for writing, English handwriting by different nationals must have the unique difference in terms of writing style, spacing between characters and words, stroke pattern, etc. This is valid because we target the people who born in the respective countries but not immigrants and non-immigrants. To extract such observation, motivated by the method [7] where Cloud of Line Distribution (COLD) has been used for writer identification, we explore the same with new feature extraction for nationality identification in this work.

II. PROPOSED METHODOLOGY

This work considers English handwritten documents of five nationals, namely, Bangladesh, China, India, Malaysia and Iran for identification as mentioned in the previous section. Since the proposed work aims at identification of nationals at text line level, we segment text lines from handwritten documents based on horizontal projection profile analysis computed using tangent angle and the mean of intensity values of each row in the image. It can be expected that persons can use ruled or unruled page for writing. When a document image contains ruler, it poses a problem for extracting text components from text line images as it connects text components of lines as one big component. Therefore, the proposed work removes ruler from images based on tangent angle between character contours and ruler. As nationality changes, writing style also changes. It is valid because every national has the own unique national script and uses the same script for writing most time.

Motivated by the method [7], which explores COLD features for writer identification, we propose the same COLD for nationality identification through handwritten text lines analysis as it helps in studying character shapes of handwritten text lines. This gives different shapes of COLD distribution for different nationals, and gives almost the same distribution for writers of respective nations. In other words, the shape of distribution shares the common properties for text lines written by different writers of the same nation. To study shape of COLD distribution, we propose a new way of feature extraction by estimating distance from pixels to the principal axis of the distribution, which gives a feature matrix. The reason to choose principal axis is that it is independent of rotation, and is

insensitive to pixels of the distribution because the principal axis is drawn based on the direction of majority pixels. That is if a distribution loses a few pixels, it does not affect for studying its shape through the principal axis. Furthermore, the feature matrix is subjected to an SVM classifier for nationality identification. The steps and flow of the proposed method are shown in Fig. 1.
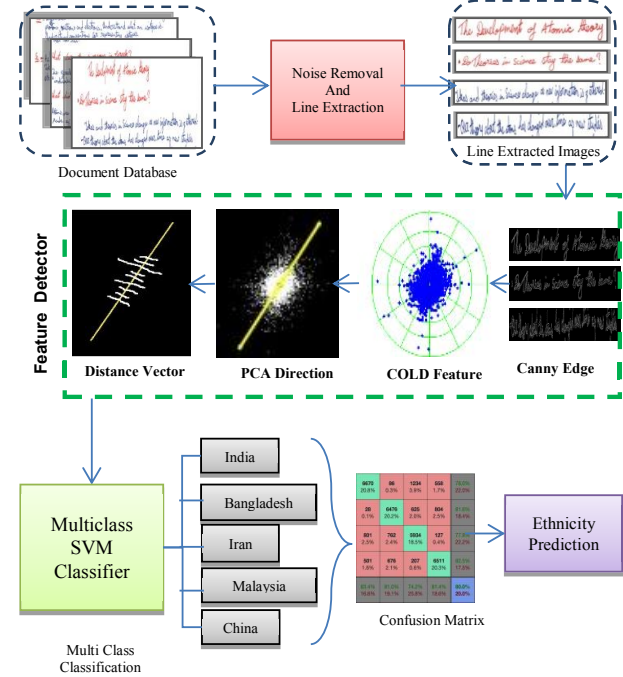

Fig 2. Framework of the proposed method

A. Preprocessing for Text Components Detection

For removing ruled lines in the input image, motivated by the method [8] where the Horizontal Projection Profile (HPP) and base line are used for bank cheque processing, we use the same for removing ruled line in this work. For the pixel values in each row of the image, the proposed method computes mean as defined in Equation (1), where $c_i(x_i, y_i)$ denote intensity values in each row and n is the number of pixels in each row. Then it estimates tangent angle for the corresponding white pixel in Canny edge image of the input text line image as defined in Equation (2), where x and y denote coordinate of current pixel and its neighbor pixels in the row.

$$HPP = \frac{\sum_{i=1}^{n} c_i(x_i, y_i)}{n} \qquad (1)$$

$$\emptyset = argtan\left(\frac{canny(y_i) - canny(y_{i+1})}{canny(x_i) - canny(x_{i+1})}\right) \qquad (2)$$

When there is a text, the mean intensity gives high values, at the same time, tangent angle give arbitrary. When there is no text, the mean intensity gives almost near to zero while tangent gives the angle as the same as the base line for border pixels. Based on these cues, the proposed method removes ruled lines from the input image as shown in Fig. 3, where in (a) we can see

524

a ruled line and in (b) the ruled line is removed. This idea also works well for text line images that do not have ruled lines. For those text line images, the tangent angle gives arbitrary angle like other edge pixels. If the text line image is affected by noises, it does not affect much for removing ruled lines because noise may not change the tangent angle of the pixels at border and the base line. In general, one can expect a horizontal handwritten text with little tilt but not with different rotations. Besides, the proposed work uses the latest Camera for capturing images, noise is not trivial in contrast to scanned images. Therefore, noise cannot be threat for the proposed work. For the output of the pre-processing steps, the proposed method obtains Canny edge images for respective five nations handwritten images as shown in Fig 4. The edge components in the images are considered for generating COLD distribution, which will be discussed in the next section.

(a) Segmented text line from the ruled Bangladesh document

(b) Rule line is removed from the image in (a).
Fig. 3. Preprocessing step for feature extraction.

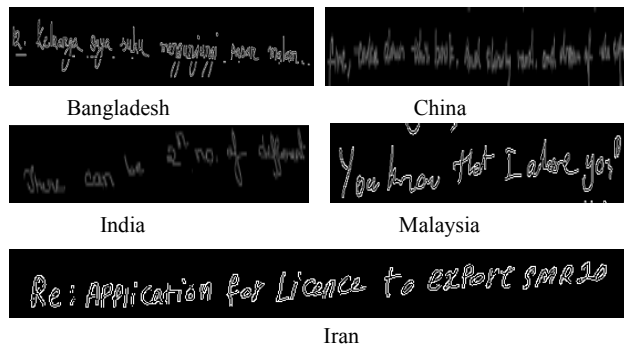| Bangladesh | China |
| India | Malaysia |

Iran
Fig. 4. Canny edge images of handwritten line images of five nations.

B. COLD Distribution for Feature Extraction

As discussed in the Introduction Section, the character shapes changes according to variations of handwriting by different nationals, which in turn changes the stroke of characters. To study such behavior of strokes, it is necessary to understand curvy and straightness of strokes. Therefore, inspired by the method [7] where Cloud of Line Distribution (COLD) is proposed for writer identification, we propose the same for extracting variations in handwriting of different nations. For each edge component in the text line images as shown in Fig. 5(a) and Fig. 5(b), the proposed method finds dominant points for the contours using polygonal approximation algorithm [7] as shown in Fig. 5(c).

Let dominant points be over contour of edge components, $D = \{ P_i(x_i, y_i) | i = 1,2,3 \ldots . N\}$ where P denotes a dominant pixel, (x_i, y_i) denotes x and y coordinates of the dominant pixel,

and N is the total number of dominant points. With these dominant points, the proposed method forms lines by joining nearest neighbor dominant points as shown in Fig. 5(c) where we can see dominant points pair, which results in line segments. To study the straightness and cursiveness of the line segments in polar coordinate domain, the proposed method estimates angle, θ for the line segments of the contour using x and y coordinates as defined in Equation (4). The length, r of line segments is determined as defined in Equation (5).

$$\theta = \tan^{-1}\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right) \tag{4}$$

$$r = abs\left(\sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2}\right) \tag{5}$$

Here (x_i, y_i) and (x_{i+1}, y_{i+1}) denote the coordinates of a dominant pair. A line segment can be represented using θ and r as a point (θ, r) in polar domain. We draw points for all the line segments in polar domain, which results in a distribution as shown in Fig. 5(d), where one can see COLD distribution for the line segments in Fig. 5(c).
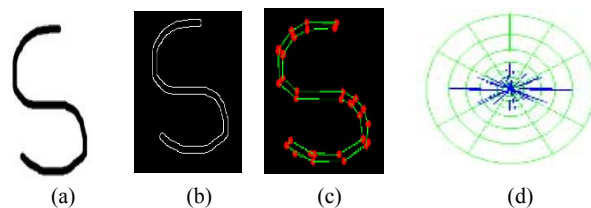

(a)   (b)   (c)   (d)
Fig. 5. COLD distribution for handwriting components. (a) is a handwritten character, (b) is the Canny edge image of (a), (c) shows the dominant points for the contours, and (d) gives Cloud of Line Distribution (COLD) in polar coordinate.

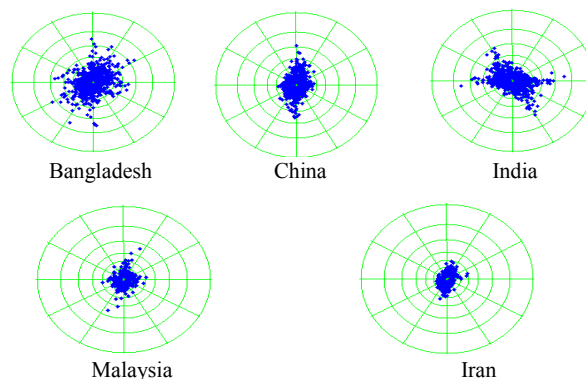| Bangladesh | China | India |

| Malaysia | Iran |
Fig. 6. COLD distribution of text lines of five nations.

The shape of COLD changes according to the behavior of line segments of character components. For example, if a character is too cursive, in this case, one should expect more short line segments. When we have more short lines, the COLD distribution appears dense at the center origin. If line segments are too long, one can expect scattered COLD distribution. In other words, if character components have regular curvature, it is expected centralized COLD distribution with a particular angle. If character components have irregular curvature, it is

525

expected scattered COLD distribution with arbitrary angle. It is evident from the COLD distribution of five nationals shown in Fig. 6, it can be seen different distributions for all the five nationals. It is observed from Fig. 6 that since handwriting of Chinese nationals involve more vertical and horizontal strokes, the COLD distribution appears dense at the center region without scattered pixels. The same thing is true for handwriting of Malaysia and Iran with tiny differences. However, for Bangladesh and India, since they write many cursive characters that may involve regular and irregular curvature, the COLD distributions appear scattered with different orientations. These observations motivated us to extract features to study shapes of distributions for nationality identification.

C. Feature Extraction for Nationality Identification

For a COLD distribution, the proposed method maps all the pixels from polar domain to spatial domain, which results in a binary image as shown in the middle column in Fig. 7. This is to reduce the implementation difficulty. For studying shape of distribution, we propose to estimate principal axis with Principal Component Analysis (PCA), which gives a direction according to majority pixel directions and hence it is invariant to rotation and scaling. In addition, if a distribution loses a few pixels, it does not affect for drawing correct principal axis according to the shape of the distribution. The proposed method considers principal axis as a reference line and moves in perpendicular direction to every pixel of reference line in both left and right directions until it reaches white pixels of the distribution in both left and right direction. When it reaches white pixels in both directions, the proposed method calculates the distance from the left pixel to the reference, that is LD, and the right pixel to the reference, that is, RD. The absolute difference of LD and RD is considered as the feature of the pair of pixels. This process continues until it finds no pixel in any of left or right direction of reference while moving in the same direction. The whole process is repeated for all the pixels of the reference line. It is illustrated in the third column in Fig. 7, where we can see the reference line is marked by yellow color and lines are drawn for estimating the distance between the left and right side pixels for COLD distributions of five nations. Finding pair pixels and absolute difference of distance help us to withstand intra variations caused by different writers of the same nation. This process outputs a feature matrix for the input text line image.

Since the proposed features have high discriminative power for classification, we propose to use well-known SVM classifier in this work rather than the recent learning tools such as deep learning. This is to show the effectiveness of the features in identifying ethnicity. For training and testing, we follow 10-fold cross validation procedure and Gaussian kernel for building SVM classifier.

III. EXPERIMENTAL RESULTS

As per our knowledge, the dataset for nationals, which involves ASIAN countries, is not available publicly. We, thus, create our own dataset by collecting English handwriting images from different ASIAN countries. There is one dataset called QUWI [9], which is available but it considers countries, which use

COLD PCA Axis Shape
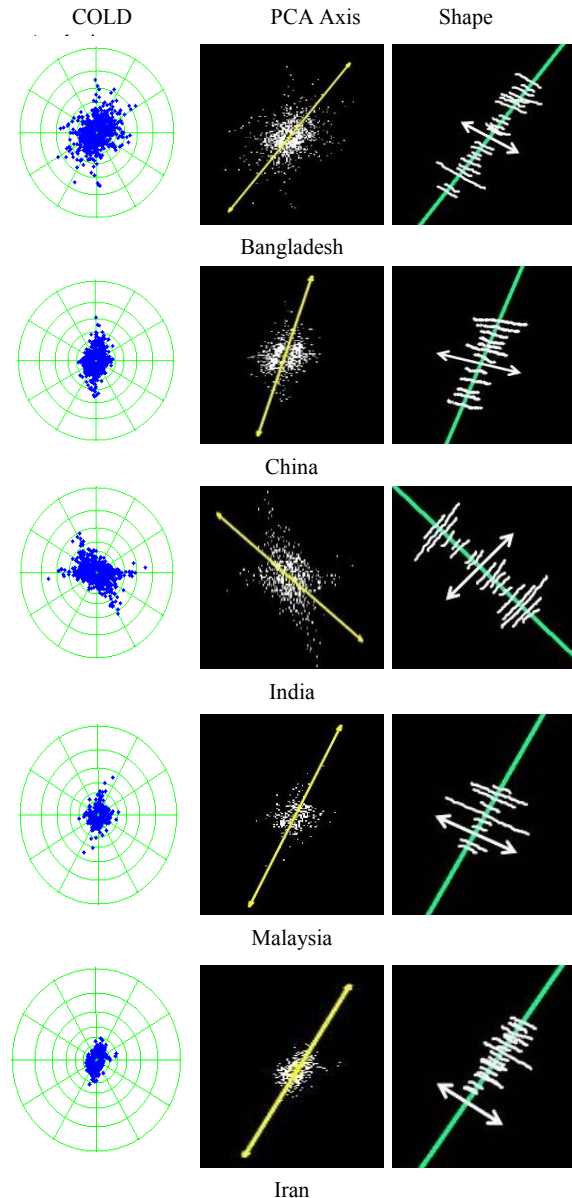
Bangladesh

China

India

Malaysia

Iran

Fig. 7. Feature extraction for handwriting text lines of five nations.

Arabic as national language and does not involve any ASIAN countries, such as India and Malaysia. Our dataset consists of English handwritten line images of different aged persons from 20-25 years who are free to use any pen, ink, paper and topic for writing. In addition, the dataset involves 100 writers for each class. Out of 100, 90% are male and 10% female, which also includes left hand and right hand writing. As a result, the dataset is considered as complex data for validating the proposed method. We consider each nation as one class and 100 text line images per class are considered. Therefore, for five nations, namely, Bangladesh, China, India, Malaysia and Iran, 500 text line images are considered for evaluating the proposed method.

526

We consider confusion matrix and Classification Rate (CR) as in the method [6] for nationality identification. To show the usefulness of the proposed method, we implement the method [6], which uses several geometric features based on contours to study the shape of character components and then identifies nationals for comparative study. However, the method is limited to nationals that consider Arabic as national language. In addition, the method reports very poor results for different features for nationality identification. This shows that the proposed geometrical features are not robust to intra and inter variations of handwritten text line images written by different nationals. The main reason to consider this method for comparative study here is that this is the state-of-the-art method for nationality identification. Though nationals are different, objective of the method is the same as the proposed method.
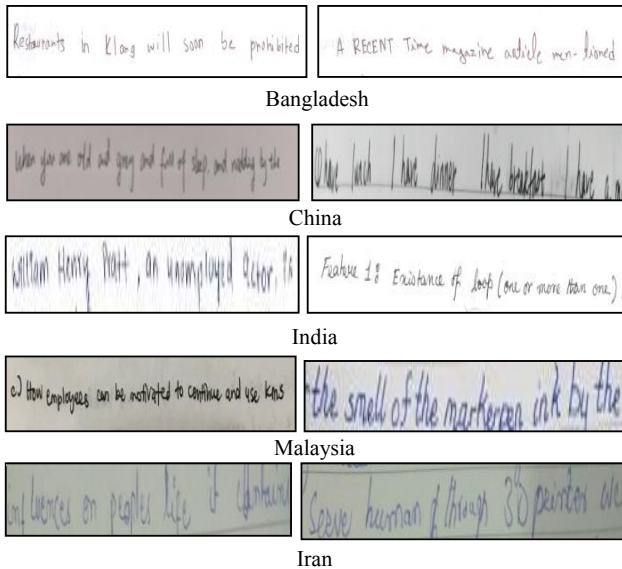

Bangladesh

China

India

Malaysia

Iran

Fig. 8. Sample successful handwritten text line images of the proposed method for five ASIAN countries.

Sample successful handwritten text line images of five nations are shown in Fig. 8, where one can notice that our dataset includes low contrast text line images affected by blur, ruled and unruled pages. Quantitative results of the proposed and existing methods are reported in Table I and Table II, respectively, where it can be noticed that the proposed method outperforms the existing method in terms of classification rate. This is valid because the proposed features especially COLD distribution is robust to intra variations of different writers of the same nation, at the same time, it provides the unique shape for handwritten text line images of different nations. In addition, the proposed features are invariant to rotation, scaling and to some extent to distortion caused by low contrast, resolution and blur. On the other hand, the proposed features in the existing method are sensitive to pixels and contours. In other words, if a character or contour loses a few pixels, the method may not work well. Therefore, the existing method reports low results compared to the proposed method.

Table I. Confusion matrix and Classification Rate (CR) of the proposed method.

Classes	Bangladesh	India	China	Iran	Malaysia
Bangladesh	**79%**	14%	0%	4%	3%
India	15%	**73%**	4%	6%	2%
China	3%	7%	**82%**	3%	5%
Iran	7%	13%	3%	**66%**	11%
Malaysia	6%	10%	7%	2%	**75%**
CR in (%)	75				

Table II. Confusion matrix and Classification Rate (CR) of the existing method [6].

Classes	Bangladesh	India	China	Iran	Malaysia
Bangladesh	**41%**	18%	10%	11%	20%
India	21%	**37%**	15%	13%	14%
China	12%	20%	**38%**	12%	18%
Iran	19%	17%	9%	**45%**	10%
Malaysia	17%	10%	23%	18%	**32%**
CR in (%)	38.6				

IV. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new method for nationality identification. It does preprocessing for removing ruled lines in handwritten document images, and explores the Cloud of Line Distribution for studying shapes of character components. The unique shapes of distributions of different nations are studied based on new features, which are subjected to an SVM classifier for nationality identification. In the future, we plan to identify provinces in multi-lingual countries.

REFERENCES

[1] M. Singh, S. Nagapal, M. Vatsa, R. Singh, A. Noore and A. Majumdar, "Gender and ethnicity classification of iris images using deep class encoder", In Proc. IJCB, 2017, pp 666-673.

[2] J. Heo and M. Savvides, "Gender and ethnicity specific generic elastic models from a single 2D image for novel 2D pose face synthesis and recognition", IEEE Trans. PAMI, Vol. 34, 2012, pp 2341-2350.

[3] M. A. Uddin and S. A. Chowdhury, "An integrated approach to classify gender and ethnicity", In Proc. ICISET, 2016, pp 1-14.

[4] M. Fairhurst, M. Erbilek and C. Li, "Study of automatic prediction of emotion from handwriting samples", IET Biom, 2015, pp 90-97.

[5] N. Bouadjenek, H. Nemmour and Y. Chibani, "Robust soft biometrics prediction from offline handwriting analysis", Applied Soft Computing, 46, 2016, pp 980-990.

[6] S. A. Maadeed and A. Hassine, "Automatic prediction of age, gender and nationality in offline handwriting", EURASIP Journal on Image Processing, 2014, pp 1-10.

[7] S. He and L. Schomaker, "Beyond OCR: Multi-faceted understanding of handwritten document characteristics", Pattern Recognition, 2017, pp 321-333.

[8] A. L. Koerich and L. L. Ling, "A system for automatic extraction of the user entered data from banckchecks", In Proc. SIBGRAPI, 1998.

[9] S. A. Maadeed, W. Ayouby, A. Hassaine and J. M. Aljaam, "QUWI: An Arabic and English handwriting dataset for offline writer identification", In Proc. ICFHR, 2012, pp 746-751.