Determination Of Writer's Age Using English Handwritten Documents

Shubhangi D.C.
Research scholar, Dr. M.G.R. Educational And Research Institute University, Chennai, India.
09448716838
shubhangidc@yahoo.co.in

Dr. P.S.Hiremath
Professor And Chairman Department Of Computer Science And Research Centre Gulbarga University, Gulbarga Karnataka, India 09480226698
pshiremath@hotmail.com

ABSTRACT

In this paper we investigate an entirely new domain and application of image processing and forensic science, i.e. detecting the human age from the handwritten documents using multi class support vector machines. There have been significant works in the direction of document analysis, character segmentation and hand writing recognition. Writing style of human being to a great extent varies as his age varies. In this paper we propose techniques that establishes a relationship between his age and his writing style and try to determine the age from the handwriting. Here novel chain code approach and A new type of generalized chain code(GCC) is proposed for handwritten data. Distinctive features for each character are then extracted based on the normalized GCC values. By using higher statistical moments for the above NGCC values feature set is obtained. Those features are passed to multiclass svm classifier which generate the hyperplane. Multiclass hyperplane plots the values of test images in the classified class. Here each classified class belongs to one age group. Total five age groups are shown . Results shows an efficiency of nearly 84.54%for age group 1, 79.52 for age group 2, 80.67 for age group 3, 82.58 for age group 4 and 90.467 for age group 5 which can be considered very high as there is no linear relationship of the document image and the age.

Categories and Subject Descriptors

1.5.2 [Pattern recognition]: Design Methodology – *Classifier design and evaluation.*

General Terms

Design , Performance.

Keywords

Age detection, multi class SVM classifier, GCC

1 INTRODUCTION

Handwriting is a personal biometric that is considered to be unique to an individual . As a result , the use of handwriting signature has been , for many centuries, a legally accepted means of authenticating various documents. In addition, in some criminal cases analysis of handwriting is often performed by forensic document examiners to determine the authorship of a questioned document. The methods used by forensic document examiners to reach their expert opinion are based on a set of techniques that are standardized and well documented in various texts [14, 26]. However, unlike other forms of forensic analysis (such as DNA testing and chemical analysis of material, blood and tissue samples) the analysis of handwriting does not have a strong underlying scientific base. In a recent case the scientific acceptability of in the United States courts [23]. There is, therefore, a need for scientific study to support the analysis methods used for the presentation of evidence in court. forensic document examination has been successfully challenged A detailed study of the possibility to identify a person by his/her handwriting has been carried out in several previous works [18, 19, 20]. A person's handwriting the script and its placing on the page express the unique impulses of the individual, logically, the brain sends signals along the muscles to the writing implement they control. S Srihari , C Huang , H Srinivasan [16, 17] studied and reported on the discriminability of handwriting of the handwriting of twins and individuality of handwriting. GX Tan, C Viard-Gaudin, AC Kot [4] proposed the an automatic text independent writer identification framework for online documents using character prototypes. By examining a handwriting sample, an expert graphologist is able to identify relevant features of the handwritten script, and the way the features interact. The features, and interaction between them, provide the information for the analysis. In the proposed method , we have shown that some computational features can be used for automatic age detection of writer to achieve high verification accuracy. Computational features of handwriting are the features that are unambiguously defined and hence can be reliable measured from images of handwriting. Vladimir pervouchin and graham ludhan[25] was shown the direct analysis of feature analysis for writer identification. Computational features may or may not be correspond to any features of forensic document examiner use in the analysis of handwriting. In the proposed method we are using chain codes, here the generalized chain code scheme was used to encode an isolated English handwritten characters and NGCC values were generated. Chain codes, pioneered by Freeman, were originally developed for data compression on line drawing and planar curves data [5 ,10], and recently extended to handwriting and line graphics [6, 7, 11]. Chain codes are used to represent a boundary by a connected sequence of straight line segment of specified length and direction. Typically , this representation is based on 4 or 8 connectivity of segments. The direction of each segment is coded by using a numbering scheme. Conventional generalized chain coding (GCC) uses a set of concentric coding rings for adaptive encoding but the coding ring set must be predetermined for efficient code assignment [10]. We extend the generalized chain coding concept to novel GCC coding scheme for recognition application. The shape of a curve can be described by the new GCC codes, which constitutes the basis for feature extraction. The writing speed information is valuable for applications such as writer recognition/ identification. The proposed new generalized chain code is able to preserve such writing characteristics. The feature set of proposed method provide the valuable information for determination of writers age.

2 NEW GENERALIZED CHAIN CODE

Hung Yuen [8] had given the chain coding approach for real time recognition of ON-LINE Handwritten characters. Chain coding is a spatially differential coding technique. To encode a curve, one must find the nearest vector node intersected by the curve on a square coding ring centred at the previous encoding point. Generalized chain codes are a special class of chain codes that uses multiple concentric coding rings. For higher efficiency, the number of ring sizes that can be used is limited, and they are usually predetermined in order to assign unique binary codes to all the vector nodes. The actual coding ring size used at each coding step is determined based on the smoothness of the trace in such a way that a longer straight curve segment should be encoded by a larger coding ring. Fig. 1 shows the conventional (1, 2, 3)-GCC coding rings.

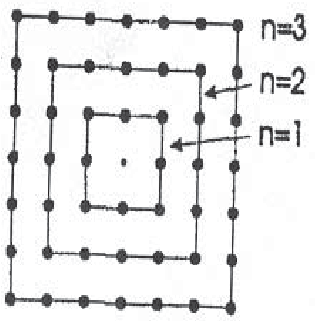


Figure 1: Conventional (1,2,3)-GCC coding rings

The proposed GCC code contains two parameters. The first parameter is the specific coding ring order ***n***. A coding ring contains M nodes, where M = 8n and ***n*** = 1, 2, 3, **....** The ring order ***n*** can be determined by the max{ x-coordinate difference, y-coordinate difference } between two successive data points. The second parameter is the actual node number ***i,*** where ***i*** = 0, 1, 2, ..., ***8n*** - 1. Hence, every handwritten data point except the starting point can be completely specified by these two parameters, based on the specific coding ring centred at the previous data point. The proposed GCC code encodes the angular and spacing information for a vector link between two successive data points. Compared to the zone coding developed for telewriting in Sketchphone [22], the proposed GCC scheme is much simpler since zone coding uses zone codes to encode differential quadrant and zone information. We define the normalized GCC (NGCC) code Ci as

Ci = i/n

The NGCC code Ci represents the absolute angular information of each GCC link, independent of the specific GCC ring order ***n*** being used. For example, if the NGCC code of a link is 1.78, this link must lie between the two links connecting the primary node 1 and 2, and its angle θ should be 45" < θ < 90". The proposed NGCC coding is equivalent to an angular quantization shown in Fig. 2

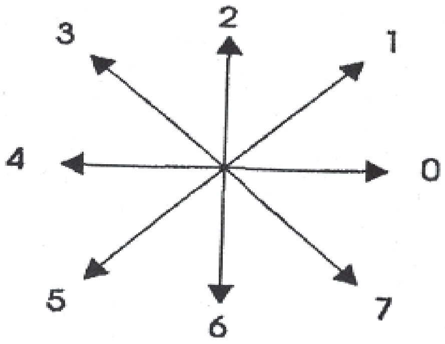


Figure 2: Angular Quantization Based On NGCC Values

Therefore the proposed GCC code preserves the dynamic writing information, which cannot be accomplished by conventional chain coding.

3 SYSTEM ARCHITECTURE

The proposed system architecture as shown in

fig 3 consists of following steps.

1. Handwritten samples both training and test samples of different age groups.
2. Pre-processing
3. GCC Encoding
4. Multi class SVM Classifier
5. Detected Age group.

The dataset used for proposed method are the English handwritten samples(All capital letters ‘A’ to ‘Z’, All small letters ‘a’ to ‘z’ and digits ‘0’ to ‘9’) of different age groups. Total 500 samples are collected from the different age group peoples. The age groups are divided into total five different age groups as follows. The first age group is from age 1 year to 20 year. The second age group is from age 21 year to 40 year. The third age group is from the age 41 year to 60 year. The fourth age group is from age 61 year to 80 year and the fifth age group is from age 81 year to 100 year. Each age group is having total 100 different English handwritten samples. Half of the samples used as test images and half of the samples used as training images. Before preprocessing the input data samples the ten fold experiment is conducted, which mix the training and test samples in every experiment. Total such ten experiments are conducted and at the end of each experiment average recognition rate of each age group was obtained and from that the overall average recognition rate was obtained .

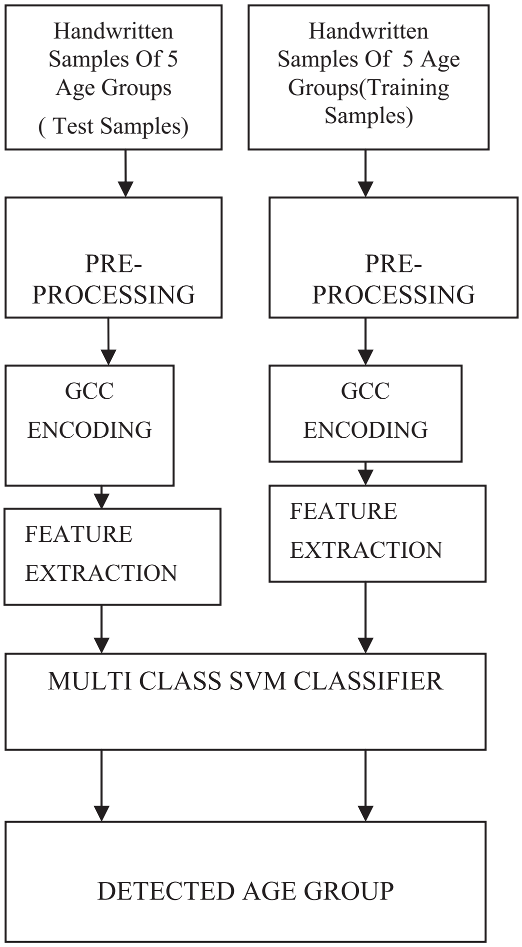

Handwritten Samples Of 5 Age Groups (Test Samples)
Handwritten Samples Of 5 Age Groups(Training Samples)
PRE-PROCESSING
PRE-PROCESSING
GCC ENCODING
GCC ENCODING
FEATURE EXTRACTION
FEATURE EXTRACTION
MULTI CLASS SVM CLASSIFIER
DETECTED AGE GROUP

Figure 3: The System Architecture

In ten fold experiment every time the training and test images are mixed and interchanged. We are calculating the average classification rate by ten fold experiment by interchanging the data in training and test dataset in every experiment by considering 50% of training and 50% of test images. The total 500 handwritten sample images are used. Each age group is having different 100 samples.

3.1 Pre-processing

The first step to any document analysis is pre-processing. First the sample database is scanned and stored as an image. This is specifically a grey scale image . A binary image is generated out of this grey scale image and one bit binary noise is removed by using erosion. The subsequent steps are all for normalizing the data to common axis.

3.2 GCC Encoding

The GCC scheme was used to encode isolated handwriting English character and NGCC values were generated for each character.

3.3 Feature Extraction

Feature extraction is an important procedure in handwriting recognition process. For recognition purpose, distinctive features that are invariant to translation, rotation and size scaling need to be extracted from the handwritten data to represent a character. The features generated in our proposed recognition method include NGCC values . NGCC values are used to describe the shape of the handwritten curves constituting a character. The NGCC codes represent the angular information of the GCC link between two successive data points. The NGCC values are generated for different handwritten characters of different age groups and the higher statistical moments are applied which generate the required feature set.

3.4 Classification

In the proposed method we are using Support vector machine classifier . The Support Vector Machine(SVM) is a new learning machine with very good generalization ability. The SVM classifier has superior recognition rates when compared to other classifiers. By applying the higher statistical moments on NGCC values for different age groups of different characters. These feature values are passed to multiclass svm classifier which generates the hyperplane which will get classified into different total five age group classes. The hyperplane provide the classified class. Each class of svm hyperplane belongs to one age group. The classified class provides the detected age groups.

4 RESULTS

Table 1 shows average age detection recognition rate of total five age groups for handwritten character samples from ‘a’ to ‘z’. Table 2 shows average age detection recognition rate of total five age groups for handwritten character samples from ‘A’ to ‘Z’.

Table 3 shows average age detection recognition rate of total five age groups for handwritten digit samples from ‘0’ to ‘9’. Table 4 shows overall average age detection recognition rate of total five age groups for all handwritten character and digit samples.

The experimental results shows that, The average recognition rate together for all handwritten English samples that is for all handwritten English capital letters and all handwritten English small letters and 0 to 9 English handwritten digits, for first age group is 84.54%. The average recognition rate for second age group is 79.52%. The average recognition rate for third age group is 80.67%. The average recognition rate for fourth age group is 82.58%. The average recognition rate for fifth age group is 90.467%. The average recognition rate of age detection is highest for the old age peoples having the age above 80 years and then the second highest recognition results are for the minor age peoples having the age below 20 years.

In the above table the following age groups are shown

Age group 1 = 1 to 20 years, Age group 2 = 21 to 40 years

Age group 3 = 41 to 60 years, Age group 4 = 61 to 80 years

Age group 5 = 81 to 100 year

Table 1. Average AGE detection recognition rate of total five age groups for handwritten samples 'a' to 'z'

H.W. Data	Avg. Age detection recognition rate of All five age groups (%)				
	Age group 1	Age group 2	Age group 3	Age group 4	Age group 5
a	74	70	78	77	82
b	84	79	74	78	81
c	76	77	79	84	85
d	82	78	84	80	86
e	79	78	81	78	80
f	84	76	79	77	87
g	88	78	84	82	90
h	89	77	80	86	91
i	84	75	78	88	94
j	88	79	81	84	92
k	79	74	76	82	88
l	81	79	77	80	84
m	86	75	79	88	94
n	91	82	88	90	98
o	85	76	77	78	91
p	93	84	83	88	90
q	77	79	84	89	93
r	87	75	82	89	92
s	85	77	78	89	94
t	92	75	78	93	98
u	79	84	83	89	93
v	91	90	80	87	94
w	77	76	80	84	89
x	87	89	90	78	93
y	76	80	84	85	92
z	83	79	86	79	98
Overall age detection %	83.73	78.5	80.88	83.92	90.34

Table 2. Average AGE detection recognition rate of total five age groups for handwritten samples 'A' to 'Z'

H.W. Data	Avg. Age detection recognition rate of All five age groups (%)				
	Age group 1	Age group 2	Age group 3	Age group 4	Age group 5
A	84	79	82	80	88
B	80	78	80	79	91
C	79	78	76	80	82
D	82	78	76	79	86
E	84	79	76	78	82
F	89	82	84	86	98
G	77	78	89	86	93
H	84	82	84	89	90
I	78	76	83	88	93
J	86	82	79	78	88
K	85	79	83	81	94
L	94	82	78	79	89
M	87	76	79	82	93
N	88	74	73	79	86
O	84	72	77	78	88
P	90	81	82	88	93
Q	89	79	86	79	89
R	77	78	74	78	83
S	93	81	78	90	94
T	84	86	82	78	91
U	75	74	80	79	93
V	92	92	79	86	95
W	98	82	87	79	97
X	84	88	90	82	94
Y	79	84	88	89	92
Z	86	82	88	83	98
Overall age detection %	84.92	80.07	81.26	82.03	90.76

Table 3. Average AGE detection recognition rate of total five age groups for handwritten English digits '0' to '9'

H.W. Digits	Avg. Age detection recognition rate of All five age groups (%)				
	Age group 1	Age group 2	Age group 3	Age group 4	Age group 5
0	84	78	79	80	92
1	83	78	74	79	81
2	82	81	77	74	88
3	78	77	74	80	86
4	79	82	81	78	89
5	93	86	79	87	91
6	87	74	78	79	90
7	89	87	79	82	93
8	94	85	81	85	94
9	88	79	84	81	96
Overall age detection %	85.7	80.7	78.6	80.5	90

Table 4. Overall Avg. Age Detection Recognition Rate Of All Five Age Groups (%) For All English Handwritten Characters And Digits .

Overall Avg. Age Detection Recognition Rate Of All Five Age Groups (%) For All English Handwritten Characters And Digits .				
Age Group 1	Age Group 2	Age Group 3	Age Group 4	Age Group 5
84.54	79.52	80.67	82.58	90.46

Table 4 shows Overall Avg. Age Detection Recognition Rate Of All Five Age Groups (%) For All English Handwritten Characters And Digits .

5 CONCLUSION

As in this paper we investigate an entirely new domain and application of image processing ,pattern recognition and forensic science. The experimental results shows that, The average recognition rate for first age group is 84.54%. The average recognition rate for second age group is 79.52%. The average recognition rate for third age group is 80.67%. The average recognition rate for fourth age group is 82.58%. The average recognition rate for fifth age group is 90.467%. Results shows good efficiency which can be considered very high as there is no linear relationship of the handwritten document image and age. The proposed GCC scheme was used to encode isolated handwritten English characters and NGCC values were generated for each character. The feature set used in proposed method provide valuable writing information , which can not be accomplished by conventional chain coding. The proposed generalized chain code is able to preserve such writing characteristics which help us to determine the writers age using English handwritten documents.

6 REFERENCES

[1] C. G. Leedham, V. Pervouchine, and W. K. Tan. Quantitative letter-level extraction and analysis of features used by document examiners. Journal of Forensic Document Examination, 16:21–40, 2004.

[2] C Huang , SN Srihari " Word Segmentation of off-line handwritten documents" in proc. Of document recognition and retrival ,2008.

[3] Cheng-Lin Liu " Handwritten Chinese Character Recognition : Effect of Shape Normalization and Feature Extraction" D.S. Doermann and S. Jaeger(Eds.) ; SACH 2006, LNCS 4768, PP. 104-128, 2008. © Springer- Verlag Berlin Heidelberg 2008.

[4] GX Tan, C Virad-Gaudin, AC Kot "Automatic writer identification framework for online handwritten documents using character prototypes" –pattern recognition , 2009 – Elsevier.

[5] H. Freeman, "Computer processing of line-drawing data," Computing Surveys, vol. 6, no. 1, pp. 57-96, 1997

[6] H. Yuen and L. Hanzo, "Adaptive fixed-length differential chain coding for transmission of line graphics," Electronics Letters, vol. 31, no. 11, pp. 862-863, May 1995.

[7] H. Yuen and L. Hanzo, "Robust differential chain coding scheme," Electronics Letters, vol. 31, no. 16.

[8] Hung Yuen A chain coding approach for real time recognition of online handwritten characters. 0-7803-3192-3/96©1996 IEEE.

[9] Imran Ahmed Siddiqui , Nicole Vincent "Writer Identification in Handwritten Documents" in proc. Of 9'th International conference on Document Analysis And Recognition (ICDAR 2007) 0-7695-2822-8/07©2007 IEEE .

[10] J.A. Saghri and H. Freeman, "Analysis of the precision of generalized chain codes for the representation of planar curves," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 3, no. 5, pp. 533-539, Sept. 1981.

[11] J.C. Arnbak, J.H. Bons, and J.W. Vieveen, "Graphical correspondence in electronic-mail networks using personal computers," IEEE Journal on Selected Areas in Communications, vol. 7, no. 2, pp. 257-267, Feb. 1989.

[12] Marius Bulaca, Lambart Schomaker " combining multiple features for text independent writer identification and verification" proc. Of tenth international workshop on frontiers in handwriting recognition (IWFHR 2006), 2006: PP. 281-286, 23-26 0ct., La Baule , France.

[13] Marius Bulacu, Lombert Schomaker " Automatic Handwriting Identification on Medieval Documents" in proc. Of 14'th international conference on image analysis and processing.(ICIAP 2007) 0-7695-2877-5107© 2007 IEEE.

[14] O. Hilton. Scientific Examination of Questioned Documents. CRC Hall, Florida, USA, 1993.

[15] P. J. Sutano, C. G. Leedham, and V. Pervouchine. Study of the consistency of some discriminatory features used by document examiners in the analysis of handwritten letter 'a' In Proc. 7th Int'l Conf. Document Analysis and Recognition (ICDAR'2003), pages 1091–1095, Edinburgh, UK, August 2003.

[16] S Srihari, C Huang, H Srinivasan " On the discriminability of the handwriting of twins" – Journal Of Forensic Sciences, 2008.

[17] S. N. Srihari , S. Cha, H. Arora, and S. Lee, "Individuality of handwriting," Journal Of Forensic Sciences, pp. 856-872 ,2002

[18] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee. Individuality of handwriting: A validation study. In Proc. 6th Int'l Conf. Document Analysis and Recognition (ICDAR'2001), pages 106–109, Seattle, USA, September 2001.

[19] [19] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee. Individuality of handwriting. Journal of Forensic Sciences, 47(4):1–17, 2002.

[20] S. N. Srihari, S.-H. Cha, and S. Lee. Establishing handwriting individuality using pattern recognition techniques. In Proc. 6th Int'l Conf. Document Analysis and Recognition (ICDAR'2001), pages 1195–1204, Seattle, USA, September 2001.

[21] SN Srihari , Catlin I. Tomai, Bin Zhang and Sangjik Leet " Individuality Of Numerals" in proc. Of seventh International Conference On Document Analysis and Recognition(ICDAR'03)

[22] T. Scheidat, C Vielhauer, J. Dittman " Handwriting verification- comparison of a multi- algorithmic and a multi-schematic approach" Image and Vision Computing, vol. 27, Issue-3, pp. 269-278 . 2'nd Feb. 2009. Elsevier.

[23] United States v. Starzecpyzel. 880 F. Supp. 1027, 1046 (S.D.N.Y. 1995), 1995.

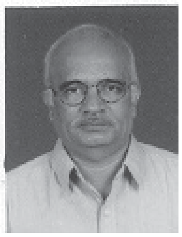
[24] V. Pervouchine, C. G. Leedham, and K. Melikhov. Handwritten character skeletonisation for forensic document analysis. In Proc. 20th Annual ACM Symposium on Applied Computing, pages 754–758, Santa Fe, NM, USA, March 2005.

[25] VLADIMIR Pervouchine and Graham Leedham Study of structural features of handwritten grapheme 'th' for writer identification. Third international symposium of information assurance and security. 0-7695-2876-7/ © 2007 IEEE, DOI 10.11.09/IAS .2007.73

[26] W. R. Harrison. Suspect Documents, Their Scientific Examinations. Nelson-Hall, Illinois, USA, 1981.

Author Biographies

Shubhangi D. C. : Received engineering degree B.E. in Electronics & communication from Marathwada university, Aurangabad in 1995, M.Tech(CSE) Degree in Visvesvaraya Technological University, Belgaum in 2000, and doing the Ph.D in computer science from Dr. M.G.R. Educational And Research Institute University Chennai. She had worked as lecturer, sr. lecturer and Asst. Professor in the Various engineering collages. She is currently working as Professor and HOD of Computer science branch of Appa Institute Of Engineering And Technology, Gulbarga. Her current Research includes pattern recognition , pattern classification and machine learning techniques. She had published five papers in International Journals and six papers in International Conferences.



Dr. P.S. Hiremath, Professor, Department of P. G. studies and Research in Computer Science, Gulbarga University, Gulbarga, Karnataka, India. He has obtained M.Sc. degree , in 1973 and Ph.D. degree in 1978 in Applied Mathematics from Karnatak University, Dharwad. He had been in the Faculty of Mathematics and Computer Science of various Institutions in India, namely, National Institute of Technology, Tiruchinapalli (1980-86), Karnatak University Dharwad (1986-1993) and has been presently working as Professor of Computer Science in Gulbarga University, Gulbarga (1993 onwards). His research areas of interest are Computational Fluid Dynamics, Optimization Techniques, Image Processing and Pattern Recognition. He has published 97 research papers in peer reviewed International Journals.