

EVALUATION OF GEOMETRIC MORPHOMETRIC APPROACH FOR ETHNICITIES DISCRIMINATION USING HANDWRITTEN NUMERAL CHARACTERS

Wan Nurul Syafawani Wan Mohd Taufek, Helmi Mohd Hadi Pritam, Wan Nur Syuhaila Mat Desa, Dzulkiflee Ismail*

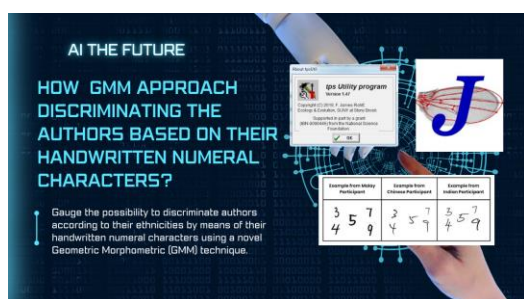
Forensic Science Programme, School of Health Sciences, Universiti Sains Malaysia, Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia

Article history

Received
18 December 2023
Received in revised form
18 February 2024
Accepted
24 April 2024
Published Online
23 June 2024

*Corresponding author
dzulkiflee@usm.my

Graphical abstract



Abstract

Handwriting evidence is a valuable source for authorship identification, an important aspect in investigating crimes such as murder, suicide, illegal drug trafficking, kidnapping, and document forgery. It relies heavily on the examination of written characters that make the document. However, specific studies on the handwritten numeral characters are scarce despite being crucial in assisting investigators in solving crimes. Hence, this study is aimed to gauge the possibility to discriminate authors according to their ethnicities by means of their handwritten numeral characters using a novel Geometric Morphometric (GMM) technique. Handwritten numeral characters collected from 30 individuals from three main different ethnic groups in Malaysia; Malay, Chinese and Indian were first digitised and landmarked using GMM software. Cluster patterns can be observed in the Principal Component Analysis (PCA) score plots, belonging exclusively to the three different ethnic groups. Significant differences ($p < 0.0001$) were discovered in handwritten numerals characters 3, 4, 5, 7 and 9 amongst the three ethnicities when tested using Procrustes ANOVA, which signifying that it is possible to discriminate authors according to their ethnicities using their handwritten numeral characters. However, more sophisticated meta-analyses are needed in order to find the most effective technique for determining and discriminating the author's ethnicity.

Keywords: Forensic science, geometric morphometric, handwritten numeral characters, handwriting, ethnicity discrimination

Abstrak

Bukti tulisan tangan merupakan bukti yang penting bagi mengenalpasti penulis, terutama dalam penyiasatan jenayah seperti pembunuhan, bunuh diri, pengedaran dadah haram, penculikan dan pemalsuan dokumen. Bukti ini bergantung ke atas pemeriksaan aksara bertulis pada dokumen yang terbabit. Walaubagaimanapun, kajian khusus atas aksara angka tulisan tangan adalah terhad, walaupun penting dalam membantu penyiasat dalam menyelesaikan sesuatu jenayah. Oleh itu, kajian ini dilaksanakan bagi bertujuan mengenalpasti kemungkinan untuk mendiskriminasi penulis mengikut etnik mereka, dengan menggunakan aksara angka tulisan tangan mereka dengan teknik Geometrik Morphometrik (GMM) novel. Aksara angka tulisan tangan dikumpul daripada 30 individu yang terdiri daripada tiga kumpulan etnik yang berbeza di Malaysia iaitu Melayu, Cina dan India, telah digitalkan dan

ditandai buat pertama kalinya dengan menggunakan perisian GMM. Corak kelompok boleh diperhatikan dalam plot skor Analisis Komponen Prinsipal (PCA), di mana kelompok tersebut milik kepada tiga kumpulan etnik yang berbeza. Perbezaan ketara ($p < 0.0001$) ditemui dalam aksara angka tulisan tangan 3, 4, 5, 7 dan 9 di antara tiga etnik setelah diuji menggunakan Procrustes ANOVA, di mana ada kemungkinan penulis dapat didiskriminasi mengikut etnik mereka dengan hanya menggunakan aksara angka tulisan tangan mereka. Walaubagaimanapun, analisis meta yang lebih canggih juga diperlukan untuk menemui teknik yang lebih berkesan bagi menentukan dan mendiskriminasi etnik penulis.

Kata kunci: Sains forensik, geometri morfometri, aksara angka bertulisan tangan, tulisan tangan, diskriminasi etnik

© 2024 Penerbit UTM Press. All rights reserved

1.0 INTRODUCTION

Numeral handwriting is a writing system that utilised a single or a combination of numeral characters to communicate numbers using writing implements. It has also been postulated that handwriting is also influenced by the ethnicity of the author [1, 2]. However, despite the fact that handwritten numeral characters can be used for identification of individuals, research involving this particular type of forensic evidence has received less attention, is insufficient and therefore becomes a challenge to the forensic document examiners (FDEs) [3]. Furthermore, the gathering of handwritten numeral characters as evidence for forensic investigation purpose should be reasonably straightforward because they are frequently encountered or being used in a variety of situations, such as postal addresses, official application forms, banker cheque, and shopping lists [4, 5].

Over the course of many years, many researchers have attempted to develop a wide range of statistical analysis with the aid of computer advancements to study handwritten numeral characters [6]. The majority of the studies are focusing on recognising the amounts or numbers written on bank cheque or postal codes on envelopes [7–9] using deep learning or neural network techniques such as Convolutional Neural Network (CNN), Multilayer Perceptron (MLP), template matching-based, K-Nearest Neighbours (K-NN) algorithms, shallow Artificial Neural Network (ANN) algorithms and Support Vector Machine (SVM). In addition, Hidden Markov Model (HMM), Deep Learning Neural Network and Discriminative Feature Extraction, Triangular Block with SVM are examples of machine learning techniques for handwritten numeral characters recognition that offer their distinctive advantages although to achieve a higher accuracy rate would require a significant effort [10]. Even though all the studies demonstrated very promising results [7–9], none of these techniques has been employed to identify authors' ethnicities, which has important bearing in forensic investigations.

As far as this study is concerned, the use of handwritten numeral characters for the discrimination of authors according to their ethnicities is limited worldwide, including in Malaysia. Malaysia is a multi-racial country which has become a melting pot of various cultural backgrounds since its declaration as an independent country. Due to the fact that criminals can come from various ethnic backgrounds, this increases the complexity for FDEs to identify and discriminate the author behind handwritten documents such as ransom notes, banker cheques, and wills.

The study involving evaluation of images of handwritten numeral characters, is an interesting field of research, howbeit it is being a challenging and time-consuming research that competes with human judgments and subjectivities [11, 12]. Another machine learning technique that has potential to be explored for handwriting analysis is the Geometric Morphometric (GMM) technique that utilised a set of pre-defined landmarks to describe, quantify and visualise the morphological shape of specimens systematically and mathematically according to their size, orientations and positions [13, 14]. Additionally, this technique can also preserve the geometrical morphological shape of the specimens and can be visualised either in two-dimensional (2D) or three-dimensional (3D) space [15]. Although GMM has been known for a while, its utilisation for handwriting analysis, let alone handwritten numeral characters for authors' ethnicities determination and discrimination is scarce.

As a result, this study, which assessed the feasibility of using the GMM technique for authors' ethnicity discrimination using handwritten numeral characters among participants from three different ethnicities in Malaysia (Malay, Chinese, and Indian) who were recruited at random at Universiti Sains Malaysia, allocating ten participants to each ethnic group, deserves forensic attention. The ethnicities selected for this preliminary study are likely to reflect Malaysia's diverse ethnic composition. Furthermore, the ethnicities chosen were consistent with the study's objectives. By incorporating the GMM approach into

this field of study, FDEs may be able to identify and discriminate against the author's ethnicity when using handwritten numeral characters as evidence in document fraud investigations involving multiple people of different ethnicities.

2.0 METHODOLOGY

2.1 Sample Collection

The handwritten numeral characters were obtained randomly from 10 Malays, 10 Chinese, and 10 Indian authors ($n=30$) at the Universiti Sains Malaysia Health Campus Kota Bharu, Kelantan, when the gender distribution was not considered as this preliminary study focused more on the ethnicity factor for this approach. The inclusion criteria included that (a) participants must be at least three generations of pure Malay, Chinese, or Indian; (b) they should be able to write and read in their native languages; (c) they should be aged between 18 and 60 years old; (d) they must be Malaysian citizens; and (e) they must be free from any hand injuries. Individuals who did not fit these criteria were not selected as participants. Each participant was asked to write numeral characters of 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 on white, unlined A4 paper using a black ballpoint pen (Faber Castell, Malaysia). The small sample size was due to the feasibility of recruiting participants at Universiti Sains Malaysia. The sample size was also constrained by time and resource limitations, which would limit the generalisability of these study findings. Therefore, this study is regarded as an initial exploration, and any interpretations drawn from it were made with caution and any generalisations avoided at our best.

2.2 Environmental Setup

An important aspect of this study was the setting up of the environment for the image's digitisation, pre-processing and data analyses. There were two different environmental setups, namely the hardware and software. For the hardware, an AMD Quad Core A12- 9720P installed Windows 10 and scanner of Epson L3250 printer (Epson, Japan) were used to digitise the handwritten numeral characters. Meanwhile for the manipulation of the scanned images, the Adobe Photoshop version 23.4, TpsUtil version 1.81, TpsDig2 version 2.31 and Notepad++ version 8.4.2 packages were used.

2.3 Data Pre-processing

Prior to performing the GMM analysis, the handwritten numeral characters obtained from all the participants were scanned using a scanner in all-in-one Epson L3250 printer (Epson, Japan). Using the PhotoScape X package, the scanned handwritten numeral character images were improved in terms of their contrast and brightness. Any extraneous pixels were

also carefully removed. The Adobe Photoshop software was used to pre-process the handwritten numeral character images while the TpsUtil software was used to convert the images from the Joint Photographic Experts Group (JPEG) into the tps. format.

2.4 GMM Analysis

The TpsDig2 software was used to digitise and landmark the coordinates on the numeral character images, while the Notepad++ was used to record and store the resultant digitised landmark coordinates generated from each of the handwritten numeral characters. The purpose of landmarking was to convert the anatomical points into shapes in the 2D dimension. The meaningful landmark points were chosen based on the class characteristics of the handwritten numeral characters, which can describe the important anatomical points geometrically [16]. The number of anatomical points selected for each handwritten numeral character varied depending on the geometry of the character itself, as long as it was sufficient to describe the morphology of numeral characters for GMM analysis. An example of points selected for each handwritten numeral character is attached in Appendix A in the supplementary section.

MorphoJ software is an integrated software package designed specifically for the geometric morphometric study and is preferred compared to other software. Morphometric analysis often uses the arrangement of morphological landmarks as the data source and extracts shape information from them by Procrustes superimposition [6]. MorphoJ software is perhaps the easiest standalone software to use by the beginner, as the graphical user interface is simple and clear, it can quickly run several analyses, and it generates fully customised graphs that can be exported as images or vectorized figures [14].

2.5 Data Analysis

Statistical analyses namely PCA and Procrustes ANOVA were carried out using the MorphoJ to generate explainable outcomes in terms of shape variations. PCA is a classical method that is commonly used to reduce the dimensionality and complexity of a multidimensional dataset by decreasing variance in order to measure the variability of shape which can lead to the identification of shape patterns and sources of variability of the shape, in our case the handwritten numeral characters [17, 18]. PCA was performed to the landmarks or the shape coordinates to allow for quantification and visualisation of the shape variation [19] within the handwritten numeral characters. Meanwhile, the Procrustes ANOVA is a quantified analysis that examines the numerical data to evaluate the significant difference between variables where p -value from the analysis was utilised [16, 20].

3.0 RESULTS AND DISCUSSION

3.1 Principal Component Analysis (PCA)

In the context of this study, we evaluated the applicability of the GMM technique which could be exploited in determining and discriminating the ethnicities of authors using handwritten numeral characters. PCA was applied to the complex dataset, aiming at reducing and condensing the dimensionality of the data while retaining the maximum amount of variance [21, 22]. This resulted in the derivation of new variables known as the principal components (PCs). These PCs are derived from linear combinations of the original variables with specific loading for each principal. Following this, the proportion of the overall shape variation described by each PC was computed from the resultant Eigenvalues, as presented in Table 1.

Table 1 Relative percentage of PCs handwritten numeral characters from 0 until 9, between Malay, Chinese and Indian authors

Handwritten numeral characters	PC	Eigenvalues	% variance	Cumulative % variance
0	1	0.0114	44.91	44.91
	2	0.0084	33.04	77.95
1	1	0.0018	68.44	68.44
	2	0.0005	18.45	86.89
2	1	0.0255	38.65	38.65
	2	0.0143	21.59	60.24
3	1	0.0221	44.04	44.04
	2	0.0114	22.69	66.73
4	1	0.0251	40.17	40.17
	2	0.0121	19.31	59.48
	3	0.0089	14.22	73.71
5	1	0.0409	61.80	61.80
	2	0.0080	12.14	73.94
6	1	0.0186	32.59	32.59
	2	0.0142	24.88	57.47
	3	0.0131	23.08	80.55
7	1	0.0178	57.59	57.59
	2	0.0084	27.03	84.62
8	1	0.0140	31.87	31.87
	2	0.0099	22.59	54.46
	3	0.0072	16.43	70.89
9	1	0.0339	54.86	54.86
	2	0.0094	15.21	70.07

In general, the first two principal components (PCs), i.e., PC1 and PC2, generated from the handwritten numeral characters of 0, 1, 2, 3, 4, 5, 7, and 9, explained more than 60% of the total variance, while the remaining PCs can be excluded because they contributed minimally to the overall variations, as can be seen in Table 1. From a geometric point of view, PC1 captured and revealed most of the variance in the dataset, followed by PC2, and so forth until a total of PCs were calculated. This was done in order to produce a straightforward result with a scatterplot while reducing the complexity of the analysis [22]. Accordingly, PC1 and PC2 for numeral

character 0 explained 77.95% of the variance. For numeral character 1, they explained 86.89% of the variance, while for numeral character 2, 60.24% of the total variance was explained by the combination of the first two PCs. Meanwhile, for numeral characters 3, 5, 7, and 9, PC1 and PC2 described 66.73%, 73.94%, 84.62%, and 70.07% of the total variance, respectively.

In our case, for the handwritten numeral characters of 4, 6 and 8, the first two PCs showed total variances that are less than 60% (59.48%, 57.47% and 54.46% respectively). Notably, this might stem from the minimal shape variations that exist within these numeral characters written by the Malay, Chinese and Indian authors. As PCA has gained widespread popularity in various applications, the aim of these PCA scatterplot images was to enhance the visual focus on the underlying data patterns and while acknowledging the trade-off in between comprehensiveness and legibility in graphical presentation, underlying mathematical principles and correlation matrix [23, 24]. To elaborate, the colour scheme associations involved blue colour which represents the Malay, red colour corresponds to the Chinese, and green signifies the Indian ethnicity which can be observed from Figure 1 to Figure 5.

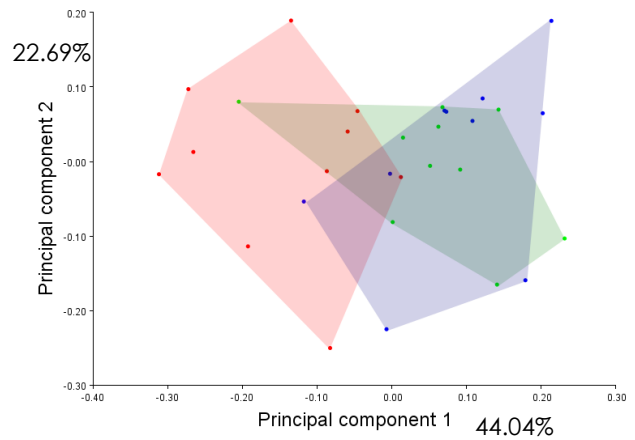


Figure 1 PCA scatterplot of the handwritten numeral character of 3 (Malay vs Chinese vs Indian authors)

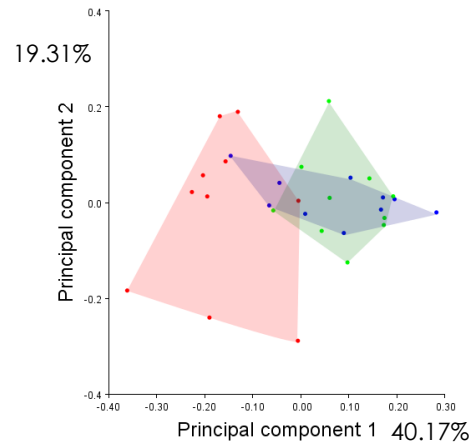


Figure 2 PCA scatterplot of the handwritten numeral character of 4 (Malay vs Chinese vs Indian authors)

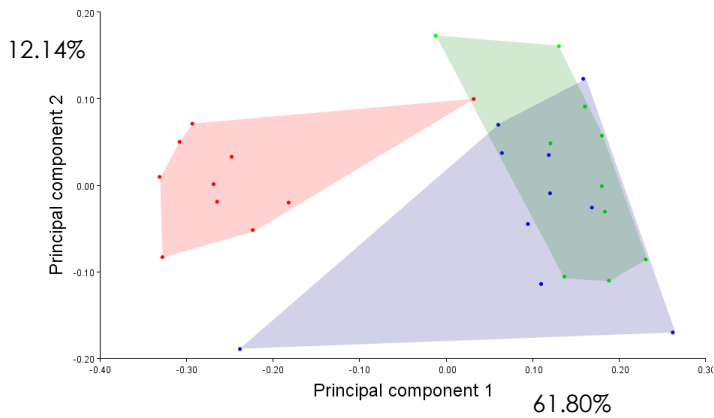


Figure 3 PCA scatterplot of the handwritten numeral character of 5 (Malay vs Chinese vs Indian authors)

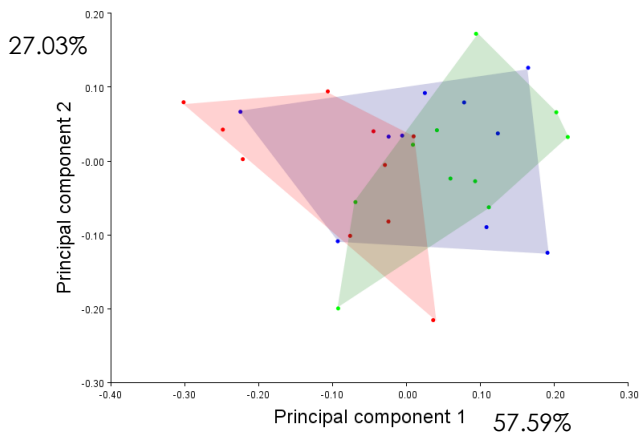


Figure 4 PCA scatterplot of the handwritten numeral character of 7 (Malay vs Chinese vs Indian authors)

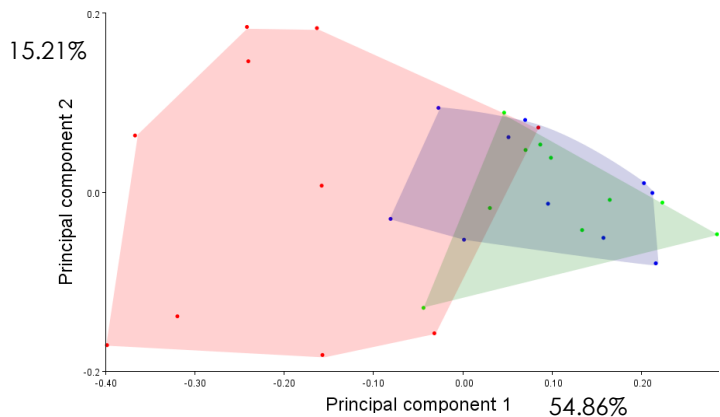


Figure 5 PCA scatterplot of the handwritten numeral character of 9 (Malay vs Chinese vs Indian authors)

Previous research conducted by Courtenay *et al.* (2018) and Nazri *et al.* (2020) have employed both 2D and 3D analyses in order to explore the shape variation based on the positioning of the landmark that marks their specimen models. Courtenay *et al.* (2018) utilised these analyses to extract features of cut marks from both 2D and 3D model specimens. On a similar vein, Nazri *et al.* (2020) employed these

analyses to assess and visually represent 50% of the pattern variance within their dataset, concentrating on the influence of sliding semi-landmarks iterations on soft-tissues based on the human facial images. Their results of the first two PCs were promising, even though they noted that some vital information might be lost due to the 2D nature of the analysis and 3D data contain more information yet less sensitive to illumination and occlusion compared to 2D analysis [24, 25].

Conversely, Otárola-Castillo *et al.* (2018) mentioned that 3D analysis is superior to 2D analysis. As 3D analysis was able to capture and differentiate the morphological variation of cutting marks, the precision of their techniques did not contrast with the 2D analysis and remained untested [26]. In almost all studies, the 3D projections of data patterns exhibit similarities to their 2D counterparts, serving to visualise the distribution of data points over a large dataset and facilitating the discrimination of clusters [27]. Nonetheless, in the context of PCA scatterplots, the utilisation of 3D data projection has been shown to improve the accuracy of pattern identification in comparison to 2D data projection [27]. A study by Tian *et al.* (2021) signified that 3D data projections through PCA scatterplots do not necessarily yield significant additional value in terms of identifying the data structures over their 2D counterparts.

We agreed that adding the additional PC not only increases the total variance, but also allows for the creation of a 3D PCA scatterplot using PC1, PC2, and PC3. This allows for a more in-depth exploration of shape variations within numeral characters; however, using only PC1 and PC2 can also result in a high level of variability, with a cumulative variance of more than 80% [28] for discriminating the authors based on their cluster patterns. Furthermore, in our case, the inclusion of PC3 does not result in a significant improvement in shape variations or author discrimination based on the resulting clusters.

PC3 can be employed in future studies, in conjunction with other multivariate statistical analyses such as Canonical Variate Analysis (CVA), Discriminant Function Analysis (DFA), K-means clustering, or Linear Discriminant Analysis, to comprehensively elucidate the shape variation and morphology of handwritten numeral characters. In addition, the use of 3D PCA scatterplots can provide valuable insights and account for 80% of the variability in the data. The benefits of 3D PCA are particularly visible when dealing with complicated data that naturally exhibits a three-dimensional structure [26]. Moreover, it is crucial to recognise that PCA is an unsupervised method that specifically aims to decrease the complexity of datasets by reducing their dimensionality. This process enhances interpretability and minimises the loss of information [29]. However, it may not inherently capture the ability to distinguish between different classes, which is essential for supervised discrimination tasks.

The study utilised PCA to decrease the dimensionality of the datasets, prioritise the variance of the data, and capture its inherent variability. PCA is useful for doing exploratory data analysis, despite its unsupervised character, which may not make it the most suitable option for discriminating analysis between the Malay, Chinese, and Indian ethnicities. Nevertheless, the issue of PCA performance can be enhanced by integrating PCA with other machine learning methods like SVM, HMM, ANN, Cloud of Line Distribution (COLD), or Random Forest-based text detection. This integration allows for the consideration of crucial information and facilitates the attainment of accurate discrimination findings.

3.2 Procrustes Analysis of Variance (ANOVA)

Table 2 represents the Procrustes ANOVA outcomes for the handwritten numeral characters of 3, 4, 5, 7, and 9, exhibiting p-values lower than 0.0001 ($p < 0.0001$). In other words, there were highly significant differences for these handwritten numeral characters written by the authors among the three different ethnicities. In contrast, the handwritten numeral characters 0, 1, 2, 6, and 8 displayed p-values of more than 0.0001 ($p = 0.5832, 0.0040, 0.0022, 0.0038, \text{ and } 0.0209$, respectively), indicating no significant differences were observed for these handwritten numeral characters written by the authors among three different ethnicities. These findings are consistent with those obtained from PCA scatterplots, as in the Procrustes ANOVA, the percentage of sum of squares can be explained identically as the percentage of total variance [30]. Unfortunately, considering the scarcity of similar studies in the literature, especially relating to multiracial populations like Malaysia, suitable comparison and discussion could not be attempted.

Table 2 Procrustes ANOVA of handwritten numeral characters between Malay, Chinese and Indian authors

Effect	SS	MS	Df	F	p-values
Shape on Numeral Characters					
0	0.0425	0.0053	8	0.82	0.5832
1	0.1231	0.0010	12	2.56	0.0040
2	0.2581	0.0108	24	2.10	0.0022
3	0.3190	0.0266	12	3.78	<0.0001
4	0.4829	0.0302	16	4.91	<0.0001
5	0.9311	0.0388	24	12.70	<0.0001
6	0.2411	0.0151	16	2.31	0.0038
7	0.1686	0.0084	20	3.12	<0.0001
8	0.1500	0.0075	20	1.80	0.0209
9	0.6262	0.0391	16	7.26	<0.0001

Note: SS: sums of squares; MS: mean squares; df: degrees of freedom; F: statistics and parametric p-values are provided for each effect.

Significant variations were observed in the handwritten numerical characters 3, 4, 5, 7, and 9 when comparing authors of Malay, Chinese, and Indian descent. While the remaining handwritten

numeral characters, such as numerical character 1, exhibited the highest cumulative percentage of variance, this does not necessarily imply that they correspond to the form variation relevant to the Procrustes ANOVA. The variation recorded by PCA may reflect the highest amount of variance that is not related to the specific morphological shape we are interested in. Therefore, Principal Component Analysis (PCA) was employed in this work to perform clustering analysis by extracting the most essential characteristics of the data. PCA facilitates the visualisation of shape variation by reducing the dimensionality of the dataset. This simplifies the identification and differentiation of patterns or clusters.

Based on Procrustes ANOVA findings, the PCA scatterplots were generated, and three distinct clusters in the data based on the PCA scatterplot on the first two PCs can be identified (Figures 1–5) for handwritten numeral characters of 3, 4, 5, 7, and 9. To further explain the total variance in these handwritten numeral characters, we have decided to exclude the PC3. This is because PCA analysis using PC1 and PC2 presents differences between the Malay, Chinese, and Indian ethnicities for handwritten numeral characters of 3, 4, 5, 7, and 9, which are strongly supported by the numeric results presented in Table 2 through a significant p-value in the case of the Procrustes ANOVA test. Roughly three clusters can be observed from Figure 1 to Figure 5, which tells us these PCA scatterplots effectively captured subtle shape variations.

3.3 The Relationship Between Handwritten Numeral Characters with The Ethnic Groups and Their Native Languages

There is distinct discrimination between at least two ethnicities, either Malay and Chinese or Indian and Chinese for handwritten numeral characters 3, 4, 5, 7 and 9, which are apparent in the PCA scatterplots. However, no distinct discrimination between Malay and Indian in any of the scatterplots which could suggest that Malay and Indian probably possess similar numeral handwritten characteristics. The shape variations of handwritten numeral characters 3, 4, 5, 7 and 9 displayed by the Chinese authors were found to have observable significance compared to those written by Malay and Indian authors. Furthermore, Chinese authors manifest distinctive clusters compared to Malay and Indian authors which are apparent in all the PCA scatterplots.

Malay language or *Bahasa Melayu* has been the common language of Malaysia, Indonesia and Brunei. In Malaysia *per se*, it is spoken daily in both official and unofficial occasions. Malay language was initially written using the Arabic characters, moving from right to left [31, 32]. However, from 1950s onwards, the Arabic characters were slowly replaced by the alphanumeric writing systems adopted from the English language. Only a handful of Malay still use the

Arabic characters particularly those who went to the Islamic religious schools whereby Tamil is the language spoken by most Malaysian Indian [31, 33].

This language is written horizontally from left to right using a set of symbols derived from the Brahmi scripts which are still practiced in India [31, 32]. Although Tamil language is still spoken by many Malaysian Indians; however, it is not the case when it comes to writing. Many Malaysian Indians are more comfortable to write using the alphanumeric writing system [31, 32, 34]. These reasons would probably explain why it is quite difficult to discriminate between Malay and Indian handwritten numeral characters.

Chinese writing system is very well known for its complexity. It does not constitute of alphanumeric but predominantly of logosyllabic, in which a character may represent a spoken Chinese word [1, 32]. Most Malaysian Chinese regardless of age at least learnt or know how to write the Chinese characters [34]. When shifting from writing Chinese characters to the alphanumeric characters, the habits of writing the Chinese characters somewhat are manifested or observable in certain alphanumeric characters. Chinese writing system contrasts with Malay and Tamil writing systems even though there is no an existence proof of direct comparison and relationship between Malay and Tamil writing systems.

This study takes its inspiration from the study by Cheng *et al.* (2005) that discussed on how native language of individuals particularly Malay, Chinese and Indian had impacted on their handwriting. The discrimination of authors using their handwritten numeral characters according to their ethnicities is possible owing to the fact that the handwriting of individuals can be influenced by the authors' habitual, native languages and writing systems [1, 16, 33]. This study has demonstrated the effectiveness of GMM as a tool to capture the 2D shape variations in the handwritten numeral characters written by Malay, Chinese and Indian authors. Based on the findings from the GMM analyses, we concluded that the handwritten numeral characters of 3, 4, 5, 7 and 9 could be potentially used to discriminate and determine the ethnicities of authors.

3.3 Application of Geometric Morphometric (GMM) for Discrimination purposes

Numerous works have explored the benefits of GMM when it comes to capturing spatial relationships more effectively. For instance, Ajanović *et al.* (2023) applied GMM in order to analyze sex discrimination of the orbital region on 211 sample of human skulls from the Bosnian population. GMM technique make it possible to quantify the shape variables of morphological structures and provides an opportunity to comprehensively analyze the overall shape of orbital region structure, regardless of curvatures and protrusions, using landmarks configuration on the examined human skulls [35]. Ajanović *et al.* (2023) concluded that sex discrimination was possible with

82.01% accuracy for males and 80.55% accuracy for females, based on the morphological variability inherent in the examined samples. However, it is noteworthy that the degree of accuracy when applying the GMM technique to a sample from another population with different geographical or demographic context might not be equivalent due to the hormonal status and differences encompassing climate, dietary habits and cultural factors [35].

Another compelling study is the work of Shin *et al.* (2021) who aimed to investigate the morphologic variations of the fourth cervical vertebrae (C4) between the different major ethnicities in the adult Malaysian population using a 3D GMM technique. By leveraging the 3D aspects of the complex datasets, Shin *et al.* (2021) achieved significantly improved determination of ethnicity accuracy compared to a traditional GMM approach. Procrustes Multivariate Analysis of Variance (MANOVA), CVA and DFA reveal statistically significant difference between the Malay, Chinese and Indian, which confirmed that there is significant difference in the variation of C4 shape between these ethnicities [36]. Shin *et al.* (2021) also mentioned that the highest discrimination accuracy was demonstrated between the Chinese and Indian, followed by between Malay and Indian, and between Malay and Chinese. Their study further revealed notable insights regarding the mean measurements of the vertebral body height, anterior-posterior length of a vertebral body, length of superior articular facet and spinous process length within C4, was the greatest for the Chinese compared to Malay and Indian, which holds potential significance for future applications, particularly in forensic contexts involving victim identification within Malaysian population [36].

Another important finding from the GMM analysis as demonstrated in this study is that the landmark configurations utilised or set to each of the handwritten numeral characters are sufficient to capture the morphological features or shape variations within and between the numeral characters written by the authors from the three different ethnicities. It is imperative to point out that although direct examinations can be used to determine the ethnicities of authors from a collection of handwritten numeral characters, this manual approach is rather laborious and time consuming. It would also require knowledge, skills, and experience on the part of the examiners. Most importantly, it is highly subjective and can result in incorrect or erroneous conclusions.

4.0 CONCLUSION

The PCA scatterplots and Procrustes ANOVA's findings from this study have demonstrated that GMM approach (an unsupervised machine learning technique) could be used to discriminate authors according to their ethnicities. GMM technique has the potential to be employed for authors' discrimination

by using their handwritten numeral characters especially in document fraud investigation where the ethnicities of authors are of concern. However, before this approach can be adopted FDEs, further study needs to be conducted. For future works, the authors plan to investigate the possibility of merging GMM technique with other methods of machine learning, such as SVM, HMM, ANN or Random Forest based text detection, in order to obtain results that are both specific and justifiable. Hopefully, this study will be able to provide FDEs with little or no experience in the GMM approach with a clear, simple and easy-to-follow step-by-step methodology which can help them to use in order to distinguish and discriminate individuals.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgements

The authors acknowledged the financial assistance given by PHA Handwriting Analysis Sdn. Bhd. and HFDE Services Ptd. Ltd (Singapore) through the industry and international grant initiatives. The authors would also like to thank all the participants and those who have been directly and indirectly involved and contributed in this study.

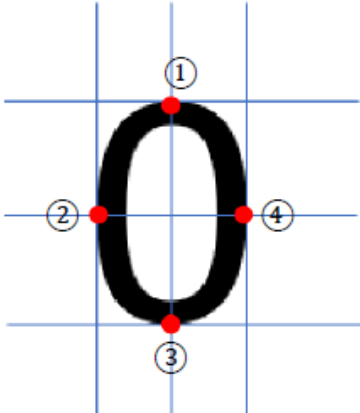
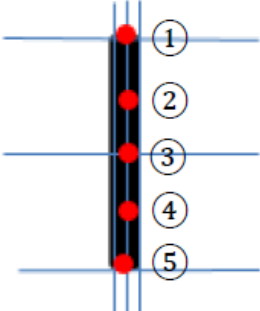
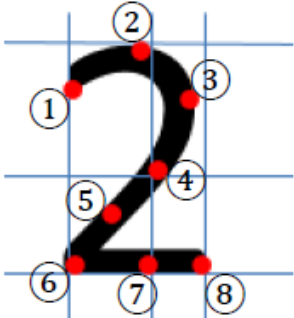
References

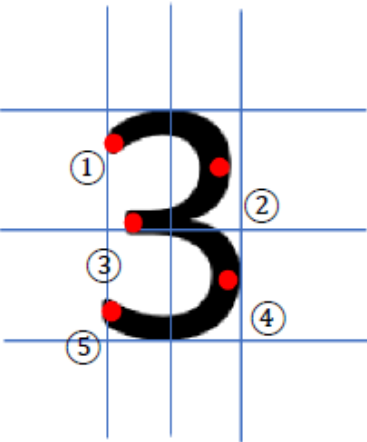
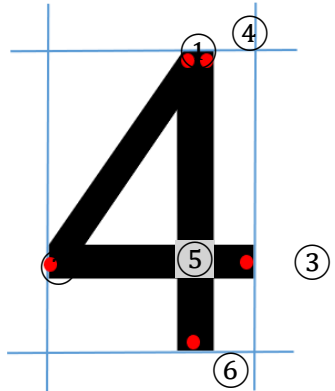
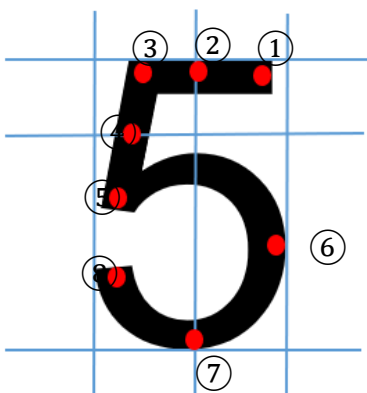
- [1] Cheng, N., Lee, G. K., Yap, B. S., Lee, L. T., Tan, S. K., & Tan, K. P. 2005. Investigation of Class Characteristics in English Handwriting of the Three Main Racial Groups: Chinese, Malay and Indian in Singapore. *Journal of Forensic Sciences*. 50(1): 1-8. <https://doi.org/10.1520/JFS2004005>.
- [2] Hame, P., Mishra, M. K., & Sodhi, G. S. 2018. Analysis of Handwriting Characteristics Based on Diverse Ethnic Distribution. *Analysis of Handwriting Characteristics based on Diverse Ethnic Distribution*. 5(2): 1-6. <https://www.researchgate.net/publication/328281692%2250>.
- [3] Deepani, V., & Kapoor, A. K. 2018. Variability in Human Handwritings: An Indian Understanding. *International Journal of Humanities and Social Sciences (IJHSS)*. 7(4): 27-32. http://www.iaset.us/view_archives.php.
- [4] Shamim, S. M., Miah, M. B. A., Sarker, A., Rana, M., & Jobair, A. Al. 2018. Handwritten Digit Recognition using Machine Learning Algorithms. *Indonesian Journal of Science and Technology*. 3(1): 29-39. <https://doi.org/10.17509/ijost.v3i1.10795>.
- [5] Bojja, P., Sai, N., Teja, S., Pandala, G. K., & Sharma, S. D. L. R. 2019. Handwritten Text Recognition using Machine Learning Techniques in Application of NLP. *International Journal of Innovative Technology and Exploring Engineering*. 9(2): 1394-1397. <https://doi.org/10.35940/ijitee.a4748.129219>.
- [6] Klingenberg, C. P. 2016. Size, Shape, and Form: Concepts of Allometry in Geometric Morphometrics. *Development Genes and Evolution*. 226(3): 113-137. <https://doi.org/10.1007/s00427-016-0539-2>
- [7] Boukharouba, A., & Bennis, A. 2017. Novel Feature Extraction Technique for the Recognition of Handwritten Digits. *Applied Computing and Informatics*. 13(1): 19-26. <https://doi.org/10.1016/j.aci.2015.05.001>.
- [8] Rao, Z., Zeng, C., Wu, M., Wang, Z., Zhao, N., Liu, M., & Wan, X. 2018. Research on A Handwritten Character Recognition Algorithm based on An Extended Nonlinear Kernel Residual Network. *KSI Transactions on Internet and Information Systems*. 12(1): 413-435. <https://doi.org/10.3837/tiis.2018.01.020>.
- [9] Biswas, A., & Islam, Md. S. 2021. An Efficient CNN Model for Automated Digital Handwritten Digit Classification. *Journal of Information Systems Engineering and Business Intelligence*. 7(1): 42. <https://doi.org/10.20473/jisebi.7.1.42-55>.
- [10] Arbain, N. A., Azmi, M. S., Muda, A. K., Muda, N. A., & Radzid, A. R. 2018. Offline Handwritten Digit Recognition using Triangle Geometry Properties. *International Journal of Computer Information Systems and Industrial Management Applications*. 10: 87-97.
- [11] Abdulrazzaq, M. B., & Saeed, J. N. 2019. A Comparison of Three Classification Algorithms for Handwritten Digit Recognition. *2019 International Conference on Advanced Science and Engineering, ICOASE 2019*. 46080: 58-63. <https://doi.org/10.1109/ICOASE.2019.8723702>.
- [12] Yahya, A. A., Tan, J., & Hu, M. 2021. A Novel Handwritten Digit Classification System based on Convolutional Neural Network Approach. *Sensors*. 21(18): 1-26. <https://doi.org/10.3390/s21186273>.
- [13] Adams, D. C., & Otárola-Castillo, E. 2013. Geomorph: An R Package for the Collection and Analysis of Geometric Morphometric Shape Data. *Methods in Ecology and Evolution*. 4(4): 393-399. <https://doi.org/10.1111/2041-210X.12035>.
- [14] Savriama, Y. 2018. A Step-by-step Guide for Geometric Morphometrics of Floral Symmetry. *Frontiers in Plant Science*. 9(October): 1-23. <https://doi.org/10.3389/fpls.2018.01433>.
- [15] Openshaw, G. H., D'Amore, D. C., Vidal-García, M., & Scott Keogh, J. 2017. Combining Geometric Morphometric Analyses of Multiple 2D Observation Views Improves Interpretation of Evolutionary Allometry and Shape Diversification in Monitor Lizard (*Varanus*) Crania. *Biological Journal of the Linnean Society*. 120(3): 539-552. <https://doi.org/10.1111/bij.12899>.
- [16] Taufek, W. N. S. W. M., Pritam, H. M. H., Desa, W. N. S. M., & Ismai, D. 2023. Identification of Writers' Ethnicity using Handwritten Numeral Characters in Combination with Novel Geometric Morphometric (GMM) Technique. *AIP Conference Proceedings*. 2896(1): 050017. <https://doi.org/10.1063/5.0177549>.
- [17] Huanca Ghislanzoni, L., Lione, R., Cozza, P., & Franchi, L. 2017. Measuring 3D Shape in Orthodontics through Geometric Morphometrics. *Progress in Orthodontics*. 18(1). <https://doi.org/10.1186/s40510-017-0194-9>.
- [18] Gorgoglione, A., Gregorio, J., Ríos, A., Alonso, J., Chreties, C., & Fossati, M. 2020. Influence of Land Use/Land Cover on Surface-water Quality of Santa Lucia River, Uruguay. *Sustainability (Switzerland)*. 12(11). <https://doi.org/10.3390/su12114692>.
- [19] Sanfilippo, P. G., Hewitt, A. W., Mountain, J. A., & Mackey, D. A. 2013. A Geometric Morphometric Assessment of Hand Shape and Comparison to the 2D: 4D Digit Ratio as a Marker of Sexual Dimorphism. *Twin Research and Human Genetics*. 16(2): 590-600. <https://doi.org/10.1017/thg.2013.5>.
- [20] Yong, R., Ranjitkar, S., Lekkas, D., Halazonetis, D., Evans, A., Brook, A., & Townsend, G. 2018. Three-dimensional (3D) Geometric Morphometric Analysis of Human Premolars to Assess Sexual Dimorphism and Biological Ancestry in Australian Populations. *American Journal of Physical Anthropology*. 166(2): 373-385. <https://doi.org/10.1002/ajpa.23438>.

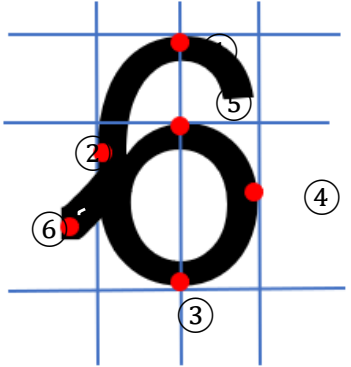
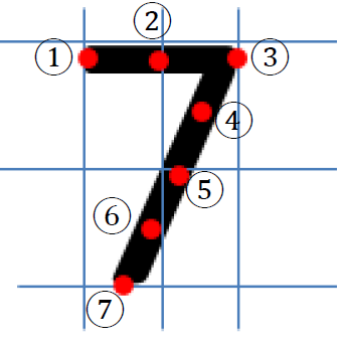
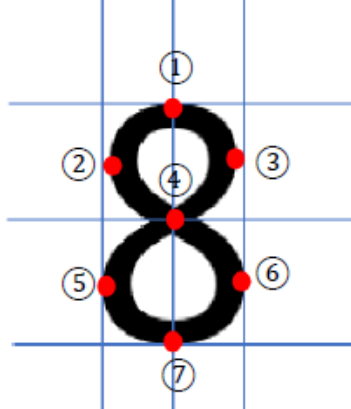
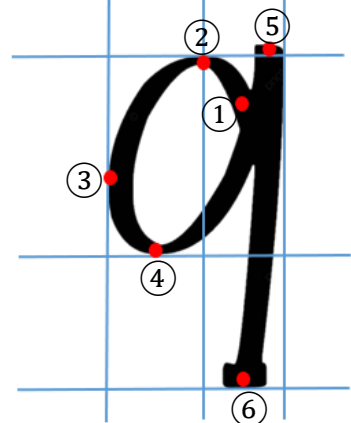
- [21] Naeim, M., Asri, M., Hashim, N. H., & Syuhaila, W. N. 2016. Pearson Product Moment Correlation (PPMC) and Principal Component Analysis (PCA) for objective comparison and source determination of unbranded black ballpoint pen inks Pearson Product Moment Correlation (PPMC) and Principal. *Australian Journal of Forensic Sciences*. 0618(November): 1-19. <https://doi.org/10.1080/00450618.2016.1236292>.
- [22] Todorov, H., Fournier, D., & Gerber, S. 2018. Principal Components Analysis: Theory and Application to Gene Expression Data Analysis. *Genomics and Computational Biology*. 4(2): 100041. <https://doi.org/10.18547/gcb.2018.vol4.iss2.e100041>.
- [23] Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D. P., Saikhom, R., Panda, S., & Laishram, M. 2017. Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). *International Journal of Livestock Research*. 7(5): 60-78. <https://doi.org/10.5455/ijlr.20170415115235>.
- [24] Courtenay, L. A., Maté-González, M. Ángel, Aramendi, J., Yravedra, J., González-Aguilera, D., & Domínguez-Rodrigo, M. 2018. Testing Accuracy in 2D and 3D Geometric Morphometric Methods for Cut Mark Identification and Classification. *PeerJ*. 2018(7). <https://doi.org/10.7717/peerj.5133>.
- [25] Nazri, A., Agbolade, O., Yaakob, R., Ghani, A. A., & Cheah, Y. K. 2020. A Novel Investigation of the Effect of Iterations in Sliding Semi-landmarks for 3D Human Facial Images. *BMC Bioinformatics*. 21(1): 1-10. <https://doi.org/10.1186/s12859-020-3497-7>.
- [26] Otárola-Castillo, E., Torquato, M. G., Hawkins, H. C., James, E., Harris, J. A., Marean, C. W., McPherron, S. P., & Thompson, J. C. 2018. Differentiating between Cutting Actions on Bone using 3D Geometric Morphometrics and Bayesian Analyses with Implications to Human Evolution. *Journal of Archaeological Science*. 89: 56-67. <https://doi.org/10.1016/j.jas.2017.10.004>.
- [27] Tian, Z., Zhai, X., van Steenpaal, G., Yu, L., Dimara, E., Espadoto, M., & Telea, A. 2021. Quantitative and Qualitative Comparison of 2d and 3d Projection Techniques for High-dimensional Data. *Information (Switzerland)*, 12(6): 1-21. <https://doi.org/10.3390/info12060239>.
- [28] Courtenay, L. A., Maté-González, M. Ángel, Aramendi, J., Yravedra, J., González-Aguilera, D., & Domínguez-Rodrigo, M. 2018. Testing Accuracy in 2D and 3D Geometric Morphometric Methods for Cut Mark Identification and Classification. *PeerJ*. 2018(7). <https://doi.org/10.7717/peerj.5133>.
- [29] Jolliffe, I. T., & Cadima, J. 2016. Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065). <https://doi.org/10.1098/rsta.2015.0202>.
- [30] Viscosi, V., & Cardini, A. 2011. Leaf Morphology, Taxonomy and Geometric Morphometrics: A Simplified Protocol for Beginners. *PLoS ONE*. 6(10). <https://doi.org/10.1371/journal.pone.0025630>.
- [31] Hoogervorst, T. G. 2015. Tracing the Linguistic Crossroads between Malay and Tamil. *Wacana*. 16(2): 249-283.
- [32] Winkler, H. 2020. Learning to Read in Multilingual Malaysia: A Focus on Bahasa Melayu, Tamil and Chinese. *GEMA Online Journal of Language Studies*. 20(1): 1-15. <https://doi.org/10.17576/gema-2020-2001-01>.
- [33] Gannetion, L., Wong, K. Y., Lim, P. Y., Chang, K. H., & Abdullah, A. F. L. 2022. An Exploratory Study on the Handwritten Allographic Features of Multi-ethnic Population with Different Educational Backgrounds. *PLoS ONE*. 17(10). <https://doi.org/10.1371/journal.pone.0268756>.
- [34] O'Brien, B. A., Mohamed, M. B. H., Yusof, N. T., & Ng, S. C. 2018. The Phonological Awareness Relation to Early Reading in English for Three Groups of Simultaneous Bilingual Children. *Reading and Writing*. 32(4): 909-937. <https://doi.org/10.1007/s11145-018-9890-1>.
- [35] Ajanović, Z., Ajanović, U., Dervišević, L., Hot, H., Voljevića, A., Talović, E., Dervišević, E., Hašimbegović, S., & Sarać-Hadžihalilović, A. 2023. A Geometric Morphometrics Approach for Sex Estimation Based on the Orbital Region of Human Skulls from Bosnian Population. *Scanning*. 2023: 1-9.
- [36] Shin, J. Y., Alias, A., Chung, E., Ng, W. L., Wu, Y. S., Gan, Q. F., Thu, K. M., & Choy, K. W. 2021. Identification of Race: A Three-dimensional Geometric Morphometric and Conventional Analysis of Human Fourth Cervical Vertebrae in Adult Malaysian Population. *Journal of Clinical and Health Sciences*. 6(1 (Special issue)): 17. [https://doi.org/10.24191/jchs.v6i1\(special\).13167](https://doi.org/10.24191/jchs.v6i1(special).13167).

Supplementary

Appendix A Definition of landmark configurations of each the handwritten numeral characters ^[16]

No.	Landmark points	Definition landmark points
1.		<ol style="list-style-type: none"> 1. The highest point of the numeral. 2. On the left side of the curve, aligned with the centre point of handwritten numeral. 3. The lowest point of the numeral. 4. On the right side of the curve, aligned with the centre point of handwritten numeral.
2.		<ol style="list-style-type: none"> 1. The highest point of the vertical stroke. 2. In between of the highest point and midpoint of the vertical stroke. 3. Midpoint of the vertical stroke. 4. In between of the midpoint and lowest point of vertical stroke. 5. The lowest point of the vertical stroke.
3.		<ol style="list-style-type: none"> 1. Initial point or stroke. 2. The highest point of the handwritten numeral. 3. On the right side, the most lateral point of the curve, where approximately $\frac{1}{4}$ of the slant line. 4. On the superior part, approximately $\frac{1}{2}$ of the slant. 5. On the inferior part, approximately $\frac{3}{4}$ of the slant. 6. On the inferior left side, the most lateral point of the handwritten numeral or horizontal stroke. 7. In between of the most inferior left lateral point and end point, on the horizontal stroke. 8. End point on the horizontal stroke.

No.	Landmark points	Definition landmark points
4.		<ol style="list-style-type: none"> 1. Initial point or stroke. 2. On the right side, superior and the most lateral point of the curve. 3. On the left side, point where the superior and inferior stroke converged. 4. On the right side, inferior and the most lateral point of the curve. 5. End point or stroke.
5.		<ol style="list-style-type: none"> 1. Initial point or stroke. 2. On the left side, most lateral point of the slant. 3. On the right side, most lateral point of the horizontal stroke. 4. The highest point of vertical stroke. 5. Intersection between horizontal and vertical strokes. 6. The lowest point or end point of the vertical stroke.
6.		<ol style="list-style-type: none"> 1. Initial point of the first stroke. 2. In between of the horizontal line or upper stroke. 3. The highest point of the vertical stroke or slant. 4. In between of the highest and lowest point of the vertical stroke. 5. The lowest point of the vertical stroke. 6. The most lateral curve on the right side. 7. In between of the most lateral curve on the right side and end point or stroke of handwritten numeral. 8. On the left side, the most lateral and inferior point of the handwritten numeral or end point or stroke.

No.	Landmark points	Definition landmark points
7.		<ol style="list-style-type: none"> 1. The highest point of the handwritten numeral. 2. On the left side, the most lateral point of the curve. 3. The lowest point of the handwritten numeral curve. 4. On the right side, most lateral point of the curve. 5. The highest point of the handwritten numeral curve. 6. End point of the handwritten numeral.
8.		<ol style="list-style-type: none"> 1. Initial point at the horizontal stroke. 2. In between starting point and the most lateral point of the handwritten numeral in the horizontal stroke. 3. Most lateral point of the handwritten numeral in the horizontal stroke. 4. On the superior part, approximately ¼ of the slant. 5. Middle between the most lateral point and the lowest point or approximately ½ of the slant. 6. On the inferior part, approximately ¾ of the slant. 7. The lowest point of the slant.
9.		<ol style="list-style-type: none"> 1. The highest point of the handwritten numeral. 2. On the left side, the most lateral point of the top circle. 3. On the right side, the most lateral point of the top circle. 4. Midpoint or Point of intersection between of the top and the bottom circle. 5. On the left side, the most lateral point of the bottom circle. 6. On the right side, the most lateral point of the bottom circle. 7. The lowest point of the handwritten numeral.
10.		<ol style="list-style-type: none"> 1. Initial point or stroke. 2. The highest point of the handwritten numeral at the curve. 3. On the left side, the most lateral point. 4. The lowest point of the handwritten numeral curve. 5. The highest point of the vertical stroke. 6. The lowest point or end point of vertical stroke.