On the Accent in Handwriting of Individuals

Faisal Farooq, Liana Lorigo, Venu Govindaraju

On the Accent in Handwriting of Individuals

Faisal Farooq, Liana Lorigo, Venu Govindaraju
CEDAR, SUNY at Buffalo
Amherst, NY 14228, USA
{ffarooq2,liana,govind}@cedar.buffalo.edu

Abstract

This paper introduces the novel concept of accent in handwritten text. Accent in speech has been studied by linguists as well as computer scientists, however, it has not been considered in handwriting. We validate this concept in handwriting. We perform our experiments with two sets of Arabic writers. The Native writers are whose first written script is Arabic and Non-Natives that wrote other scripts before learning Arabic as an additional script. Outputs from directional Gabor filters were used as input features to an SVM for classification. We achieved an overall accuracy of 86.96% *and* 78.54% *in two different settings at classifying a given word as being from either class. A perfect classification at a document level was achieved for our dataset using a max-vote scheme.*

Keywords: Accent, Handwriting, Gabor filters, SVM

1. Introduction

Accent is defined by the Webster's dictionary as an individual's distinctive or characteristic inflection, tone, or choice of words. As an alternative definition it is also described as a way of speaking typical of a particular group of people and especially of the natives or residents of a region. Examples in spoken English language are distinctive Southeast Asian (Indian), Chinese and Middle Eastern accents. As an example the various differentiating syllables in Indian accented speech from native speakers in America are a rhotic *'r'*, a rolling *'l'* and accented *'a'* [1]. In addition the accents in British English differ a lot from American English e.g. most of the speakers from England pronounce the 'a' in *path* as *'aa'* where as the American counterparts pronounce it as in the word *'act'*.

In this paper we show that accent in handwriting also exists and can help differentiate a native writer from a non-native one. To the best of our knowledge, this concept is novel in the fields of forensics and document analysis. Whereas, it could help in advancing the state-of-art in handwriting recognition (HR) by modeling the native vs non-native script features, it could also be beneficial in forensics and biometrics. It is believed that handwriting analysis could be used by forensic experts to identify an individual [11]. We hypothesize that native-script accent is a *'soft'* biometric that could aid identification of the *'group'* but could not conclusively identify an individual.

2. Previous Work

Prior work in accent classification and identification has been limited to speech where accent can be observed widespread. Researchers have studied the native vs. non-native accents as well as effect of native language of an individual on a secondary language. The tasks have thus included classification as well as identification of the accent. A study of various features that have been used to classify accent in speech can be found in [2]. Classification methods like Gaussian Mixture Models have been used in [4]. Support vector Machines and Hidden Markov Models for classification have been utilized by [12]. To our knowledge, there has been no work done on this area in the handwriting domain. There has, however, been some work on classification of handwriting into binary demographic categories like gender (male and female), age group (below 24 and above 45) or handedness (right and left) of the writer. Macro features like slant, word gap etc. and combinations of neural network classifiers based on boosting and bagging were used [3].

3. Problem Formulation

Figure 1 shows samples of a word written by six different writers. On the left are three samples from individuals that are Native Arabic writers whereas the right ones are three samples from Non-Native ones.

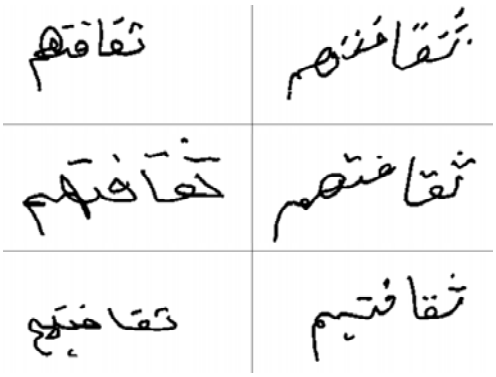


Figure 1. Handwritten samples of an Arabic word from Native (left) and Non-Native (right) individuals.

The shapes of the characters, sharpness and smoothness of curves and positioning of dots and diacritics are distinctive features amongst the two groups. This could

be attributed to the underdeveloped penmanship of non-native writers. The strokes in one script differ greatly from another. This effect is magnified when the script writing changes direction e.g. in case of Arabic (*right-to-left*) and English (*left-to-right*) (for purposes of this paper we will use English as a representative of a script).

The formation of rudimentary elements or primitives (loops, cusps, dots etc.) of the characters and the fashion in which the characters fuse are highly dependent on the motion-coordination of the hand and brain. In case of Native writers certain movements are more developed which suit the needs of that script. These movements are restricted in writers of another script that does not require their development. We believe this can be attributed to the adaptation of the muscles and the motor neurons.

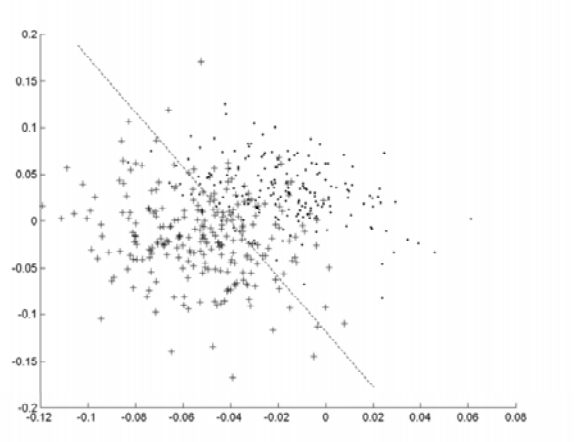


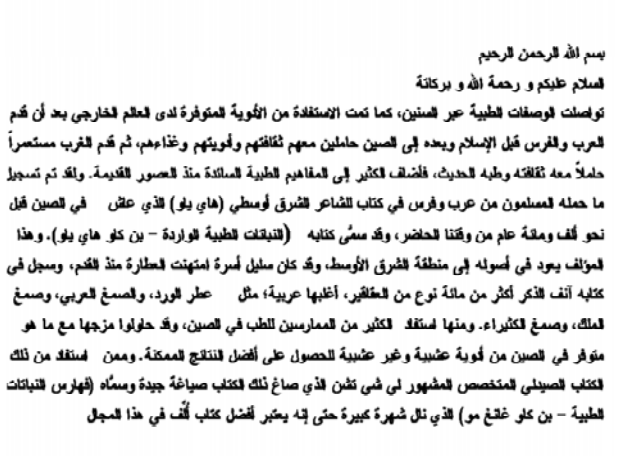
Figure 2. Principal Component Projection in $2-D$ (Native-(.), Non-natives(+)).

Based on these observations we decided to study the difference in the nature of handwriting for Native and Non-Native writers of Arabic. Figure 2 shows that the 2-D principal component analysis (PCA) scatter plot of the words in our dataset. The plot shows the projections of the two principal eigen-vectors of the features. The feature extraction is described in Section 4.2. The dots correspond to the native writers and the pluses to the non-native ones. As shown, the imaginary hyper-surface (hand-drawn diagonal line) could very well act as the discrimination boundary. The classes seem to cluster on either side of the hyper-surface. Thus, the problem was formulated as a binary classification problem - Native and Non-Native. Thus, given a document, the writer could be one of the two classes. Since the differences observed were in the writing and fusing of characters, choice of words as the unit of classification was prudent. This choice is beneficial even if full documents by a writer are not available.

4. Experimental Setup

4.1. Dataset

We collected handwriting samples from individuals whose native script was Arabic and also from those that learned to write Arabic as a second script in later stages of their lives. Thus, we had two classes - Native and Non-Native. While collecting these samples, we made no distinction about the first script of the non-native writer. Based on our experience with the Arabic script [5, 6, 9], a single page test document (see Figure 3) of ~ 200 words was created. The document contained all characters of Arabic script in almost all forms (initial, middle, final and isolated) in addition to a few ligatures e.g. *lam-alif* and diacritics. The writers were student volunteers from the University at Buffalo. We collected $\sim 3k$ words from 16 writers with an equal number of writers representing each class.


بسم الله الرحمن الرحيم

السلام عليكم و رحمة الله و بركاتة

تواصلت الوصفات الطبية عبر السنين، كما تمت الاستفادة من الأدوية المتوفرة لدى العالم الخارجي بعد أن قدم العرب والفرس قبل الإسلام وبعده إلى الصين حاملين معهم ثقافتهم وأدويتهم وغذاءهم، ثم قدم الغرب مستعيراً حاملاً معه ثقافته وطبه الحديث، فأضاف الكثير إلى المفاهيم الطبية السائدة منذ العصور القديمة. ولقد تم تسجيل ما حمله المسلمون من عرب وفرس في كتاب للشاعر الشرق أوسطي (هاي بلو) الذي عاش في الصين قبل نحو ألف ومائة عام من وقتنا الحاضر، وقد سمّى كتابه (النباتات الطبية الواردة – بن كاو هاي بلو). وهذا المؤلف يعود في أصوله إلى منطقة الشرق الأوسط، وقد كان سليل أسرة امتهنت العطارة منذ القدم، وسجل في كتابه آنف الذكر أكثر من مائة نوع من العقاقير، أغلبها عربية؛ مثل عطر الورد، والصمغ العربي، وصمغ الملك، وصمغ الكثيراء. ومنها استفاد الكثير من الممارسين للطب في الصين، وقد حاولوا مزجها مع ما هو متوفر في الصين من أدوية عشبية وغير عشبية للحصول على أفضل النتائج الممكنة. وممن استفاد من ذلك الكتاب الصيدلي المتخصص المشهور لي شي تشن الذي صاغ ذلك الكتاب صياغة جيدة وسمّاه (فهارس النباتات الطبية – بن كاو غانغ مو) الذي نال شهرة كبيرة حتى إنه يعتبر أفضل كتاب ألّف في هذا المجال

Figure 3. Original text used for collection of handwriting samples.

4.2. Feature Extraction

Gabor filters are directional filters that have been used successfully for classification of textures and automatic script identification [10]. Research has shown that a simple model for the responses of simple cells in the primary human visual cortex is the linear receptive field that can be easily modelled by a 2-D Gabor filter [8]. As examples in Figure 1 suggest, the accent in the script is a difference in the curvature and smoothness in the flow of strokes. Thus, curvature and texture that the human eye uses to differentiate between the words could be used as features for classification. Since direction of strokes and curvature is a key feature, the use of Gabor filters seem to be ideally suited for the task, as outputs of a Gabor filter yield directional strength .

Gabor functions are Gaussian functions modulated by a complex sinusoid. In $2D$, a Gabor function is given by:

$$h(x,y) = g(x',y').e^{2\pi jFx'}$$

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}[(\frac{x}{\sigma_x})^2+(\frac{y}{\sigma_y})^2]}$$

where (x',y') are rotated components of (x,y),

$$x' = xcos\theta + ysin\theta$$

$$y' = -xsin\theta + ycos\theta$$

and F is the radial frequency which for a given scale s is given by $F = F_0/s$. The output of the filter,

$$G_{\theta,s}(x,y) = \int I(s,t)h_{\theta,s}(x-s, y-t)dsdt$$

is an image with the components in the chosen direction becoming prominent.

Figure 4 shows a sample word image extracted. Figure 4 also shows the output of the Gabor filter for each direction at a single scale when applied to the word image.

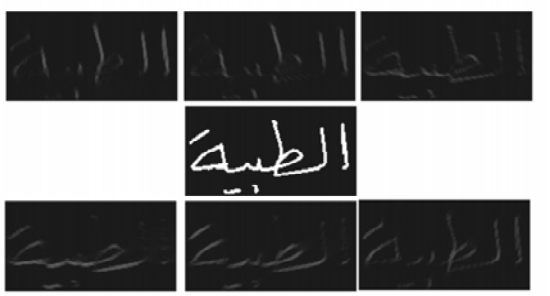


Figure 4. Gabor filter outputs for one scale and 6 directions (*center:* original word).

Since the word images all vary in their width the Gabor filter cannot be applied directly. For classifiers like neural networks or support vector machines (SVM) the feature vectors need to be of fixed size. This problem can be resolved by noting that the main information obtained from the Gabor filter output is the strength of the word image in each direction and scale which is given by the sum of the output of the filter for each direction and scale resulting in a vector of size [number of scales $\times$ number of directions]. We normalize the output by dividing the sum of filter output by the sum of the output of an isotropic Gaussian filter,

$$Gauss(x,y) = \int I(s,t)g(x-s, y-t)dsdt$$

For direction θ and scale s

$$Gabor(\theta, s) = \frac{\int G_{\theta,s}(x,y)dxdy}{\int Gauss(x,y)dxdy}$$

In our implementation we use a set of 24 filter banks with 12 directions and 2 scales. Thus for each word image we extract a 24–dimensional feature vector for classification.

4.3. Classification

Gaussian Mixture Models [4], Support vector Machines (SVM) and Hidden Markov Models(HMM) [12] have been employed for classification of speech accent. Recently Support Vector Machines have been used more commonly and are considered state-of-art, typically for a 2-class problem. Support Vector Machines are an attractive approach to data modelling. Our task involved classification into native and non-native classes and the scatter plot in Figure 2 suggests that any good binary classifier would suffice. These reasons encouraged us to use an SVM for classification. An off-the-shelf two-class SVM with a Gaussian kernel [7] was used to perform classification.

5. Experiments and Results

In order to test our hypothesis of accent in script, we performed three sets of experiments. The experiments were performed at the word level. Words were extracted from the document [5] automatically and segmentation errors corrected manually. The experiments are described in the following sections.

5.1. Random Sampling with Blind Labels

The words were labeled as classes C_1 and C_2 without the knowledge of which class was the native and non-native. We used 25% of the words for training from each class and performed testing on the remaining words. The words were drawn randomly from the uniform interval $(1, N_w)$ where N_w is the total number of words in a class. The experiment was repeated a total of 10 thousand times each time choosing a different sample set. Using this strategy the raw average accuracy achieved was 86.96% with an average percentage error of 3.91×10^{-4}. The purpose of this experiment is to show the seperability of the classes and the results were indeed encouraging. However, due to random sampling, we could use words from the same author for training as well as testing leading to the necessity of the following experiment.

5.2. Controlled Sampling with Known Labels

The second set of experiment we performed was by training the SVM with features extracted from words of only a fixed number of documents. This could help us understand if learning the general features is possible from a subset of writers in a class. In order to do this we assigned the correct labels to the words {native,non-native} instead of $\{C_1, C_2\}$. We used words from 50% of the documents for training from each class and the rest as testing. Thus, words used for testing were from a writer not used in training. We measured the performance of our system by the precision and recall metrics, commonly used by the Information Retrieval (IR) community. Precision in our case would be the ratio of words from Native labeled correctly to all words that are labeled as Native by our system. Recall is measured as the ratio of words labeled as Native correctly to all words with label Native in the test set. Similarly the corresponding metrics for Non-native are also calculated. The overall accuracy of the system (ratio of correct classifications to the size of dataset) was observed to be 78.54%. Table 1 tabulates the results. The results suggest that even simple features like the ones used have statistically significant discriminatory power. However, we believe that these results can further be improved by using more features.

Table 1. Performance analysis of the system.

	Precision(%)	Recall(%)
$Native$	69.97	86.54
$Non-Native$	89.93	76.22
Overall Accuracy(%)	78.54	

5.3. Document Class Labeling

We also performed another experiment to classify a document as belonging to a native or non-native writer. We tested all words from a document (same writer) using the SVM trained in Section 5.2. This was followed by a max-vote scheme to determine the class of the document. Thus, if the majority of words in a document were classified as native, then the author of that document was labeled as native and vice-versa. A perfect classification was achieved using this technique for documents in our dataset. The limitation for this experiment, however, is the considerably small number of the writers in our dataset.

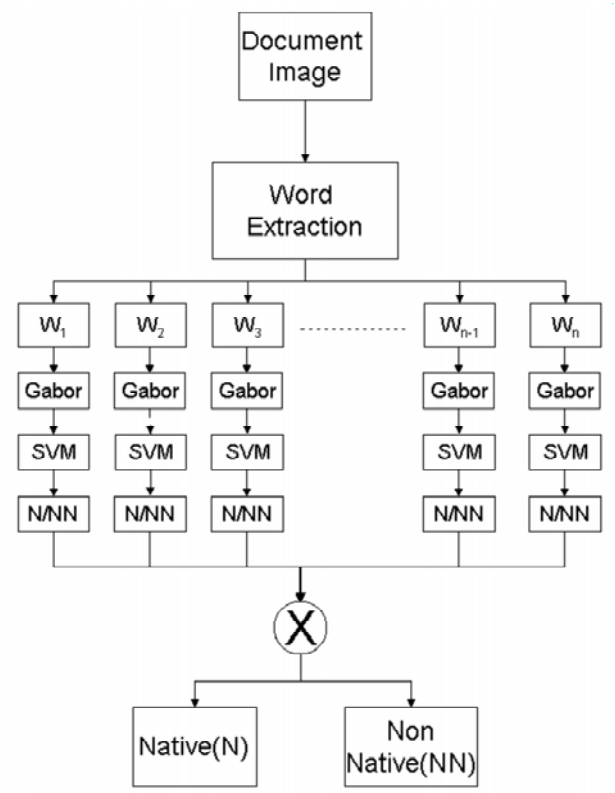


Figure 5. Classification of a given document as being from a Native (N) or Non-Native (NN) writer.

Figure 5 is a pictorial representation of the methodology where the $\otimes$ is the max operator. This was repeated for every document in the dataset.

6. Conclusion and Future Work

We have introduced a novel concept of accent in handwriting that has not been studied before by researchers. In our experiments we showed how certain characteristics exist in Native vs Non-Native handwriting. We have achieved highly encouraging results using simple features and classification schemes. In future, certain global features like loops, connectedness and curves could be used for classification. This would, of course, require more research and data. Although our experiments were performed on small datasets of Native and Non-Native writers of Arabic script, we believe that the concept is general.

An interesting direction in the future would be accent identification i.e. the identification of the accent of the Native script on another script. Whereas in our experimental setup the classes were Native Arabic writers and all writers with Arabic as their second script were classified into one group, it would be valuable to study the effect of the native script such as Arabic on the second script e.g. Latin. Specifically, given an image of handwriting in one script, we should be able to study features indicating the presence or absence of the accent of a certain script in it. For example, given a set of documents written in English, this would enable us to classify the writers that have an Arabic, Indian or Chinese accent in their English handwriting.

7. Acknowledgements

We acknowledge students of University at Buffalo for the help with data collection. We highly acknowledge Karthik Sridharan for help with the SVM code.

References

[1] *Definition of non-native pronounciations of English*. www.wordIQ.com.

[2] L. M. Arslan and J. H. L. Hansen. A study of temporal features and frequency characteristics in american foreign accent. *Journal of Acoustical Society of America*, July 1997.

[3] K. Bandi and S. N. Srihari. Writer demographic classification using bagging and boosting. *Proceedings of the 12th Conference of the International Graphonomics Society*, pages 133–137, June 2005.

[4] T. Chen, C. Huan, E. Chang, and J. Wan. Automatic accent identification using gaussian mixture model. *IEEE workshop on ASRU*, 2001.

[5] F. Farooq, V. Govindaraju, and M. Perrone. Pre-processing methods for arabic handwritten documents. *Proceedings of the International Conference on Document Analysis and Recognition*, pages 267–271, 2005.

[6] F. Farooq, V. Govindaraju, and M. Perrone. Processing of handwritten arabic documents. *Proceedings of the 12th Conference of the International Graphonomics Society*, pages 183–186, 2005.

[7] S. R. Gunn. Support vector machines for classification and regression. *Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton*, 1997.

[8] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[9] L. Lorigo and V. Govindaraju. Offline arabic handwriting recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006 *To Appear*.

[10] P. B. Pati, S. S. Raju, N. Pati, and A. G. Ramakrishnan. Gabor filters for document analysis in indian bilingual documents. *Proceedings of International Conference on Intelligent Sensing and Information Processing*, pages 123–126, 2004.

[11] S. N. Srihari, S. H. Cha, H. Arora, and S. Lee. Individuality in handwriting. *Journal of Forensic Sciences*, 47:1–17, 2002.

[12] H. Tang and A. Ghorbani. Accent classification using support vector machine and hidden markov model. *Canadian Conference on AI*, 2003.