Recommendation Systems in Mathematical Character Recognition

Vadim Mazalov and Stephen M. Watt

Department of Computer Science
The University of Western Ontario
London Ontario, Canada N6A 5B7
{vmazalov,Stephen.Watt}@uwo.ca

Abstract. In handwritten text there are usually several accepted styles for forming each character. We hypothesize that in the handwriting of individuals there is a correlation among the styles used for characters, and that these correlations may be used to anticipate which styles particular writers will use for symbols that have not yet been seen. This approach may prove useful in the setting of mathematical handwriting recognition, where there are many symbols and it would be onerous to require writers to provide samples of every one in order to personalize handwriting recognition. We describe preliminary experiments using ideas from the area of recommendation systems to predict which styles writers will likely use for symbols they have not yet written. The experiments demonstrate that writers tend to use only a small fraction of the possible combinations of character writing styles, and there are correlations among the styles used for symbols.

Keywords: Mathematical handwriting recognition, Recommendation systems, Character classification

1 Introduction

Writing style has long been taken to be a personal characteristic of an individual. Certain specific forms, such as signatures, have been used as a primary form of authentication for centuries. Conversely, writing style has also been used to narrow or even determine document authorship, when the writer is not known. We also observe that the general shape of handwritten characters may look similar among groups of individuals, especially those that have similar background, e.g. locale or period of education. We are interested in online recognition of handwritten mathematics and are currently working on improving recognition of individual characters. Earlier, we developed a cloud-based handwriting recognition framework that allows a user to share training data among devices [5]. As a side benefit to the developers, it facilitates access to the extensive amount of training data that can be indexed by different characteristics of the writer. Each new user is assigned a default training dataset. The dataset contains samples that represent different character styles (to be defined later) of the same symbol, some of which are likely to be similar to the handwriting of the new user.

However, the samples that represent character styles different from those of the new user make the training dataset noisy and may cause misclassification.

In our approach to classification, a character is represented by the coefficients of an approximation of trace curves with orthogonal polynomials [3]. Recognition is based on computation of the distance to convex hulls of nearest neighbours in the space of coefficients of approximation of symbol strokes. Typically, the method does not require many training samples to discriminate a class. However, because there is a large number of classes in handwritten mathematics, the training dataset may contain tens of thousands of characters. Therefore, any form of automated or semi-automated training can be a valuable asset in this environment.

We are motivated by the wide and successful usage of recommendation systems on the Internet that are designed to recommend products to consumers, based on their purchasing history and the history of individuals with similar behaviour [1]. In this work, we investigate similarity of character styles with respect to the writers who provided them and similarity of writers with respect to their styles. We also develop a method for semi-automated training of the recognizer by proposing character styles that are likely to be applicable to the handwriting of the new user, based on the styles that the user has already provided and the styles of writers with similar handwriting. This theory is based on the assumption that if a group of users writes some characters in the same style, it is likely that they will write certain other characters in the same style as well. An example is shown in Figure 1. This assumption is supported by an experiment we sketch in this paper.

The remainder of the article is organized as follows. In Section 2 we define some basic concepts and explain the organization of test dataset. Section 3 describes the types of handwriting similarity in which we have interest, and how we might use this to predict character styles. Section 4 presents the experimental evaluation. Section 5 gives an example of how this information can be used. Section 6 gives some conclusions.

2 Definitions and Organization of Data

In discussing similarity of handwriting we need to distinguish between various notions such as the similarity of individual symbols versus entire writing repertoires. We therefore introduce a few definitions to ensure clarity:
A *character* or *symbol* or *class* represents a single- or multi-stroke handwritten letter that may include an accent, e.g. “a”, “1”, “Σ”, etc.
A *style* or *character style* refers to the way in which one character is written. For our purposes, this is given by the class and the direction and order of the strokes in which the sample has been written. Theoretically, the number of possible styles for a single class character of k strokes is $2^k k!$, while in practice this number is not more than 3, even for samples with relatively large number of stokes.
A *writing style* is a collection of character styles for a set of characters. It may be viewed as a set of (character, character style) pairs. We may refer to an *author's*

Fig. 1. An example of characters written in a similar style (a) "9" and "a" are written clockwise, and (b) "a" and "9" are written counterclockwise

Fig. 2. (a) The structure of the dataset, (b) The structure of the user profile.

writing style to mean all the character styles observed from that author. This definition is similar to the concept of *handwriting style* investigated in [2].
A *sample* is a handwritten sample of one character provided by a user (test sample) or available in the dataset (training sample).

The dataset for our experiments has the following structure: There is an alphabet of characters C with each character $C_i \in C$ having a set S_i of corresponding character styles, as shown in Figure 2(a). There is also a set of users U. For each user $U^j \in U$ there is a set of characters $C^j \subset C$ of interest to that user. For each character $C_k^j \in C^j$ there is a style $S_k^j \in S_k$ from the set of styles with which the user writes this symbol. Each character style represents a collection of samples – the actual handwritten symbols from the user input, Figure 2(b).

3 User-Style Similarity and Character Style Prediction

Collaborative filtering recommendation algorithms are typically divided in two categories, as described in [6]. These are the item-based and user-based algorithms. Similarly, we investigate character style and writer similarity in our dataset. Further, we propose a method for prediction of character styles that are likely to be applicable to the writer.

Style-Based Similarity We propose the following measure to estimate the similarity of character styles. Consider two styles S_i and S_j, $i \neq j$ and the styles belong to classes C_i and C_j respectively. Then the style-based similarity between S_i and S_j is computed as the ratio of the number of authors who have written the class C_i and C_j respectively in styles S_i and S_j to the total number of writers who provided samples for classes C_i and C_j. This may be computed as shown in Algorithm 1.

Algorithm 1 StyleSimilarity()

Input: S_i,S_j – character styles of which to compute similarity
Output: the similarity measure

$A_i \leftarrow$ list of authors who wrote character C_i in style S_i.
$A_j \leftarrow$ list of authors who wrote character C_j in style S_j.
$A_i^0 \leftarrow$ list of authors who wrote character C_i in any style.
$A_j^0 \leftarrow$ list of authors who wrote character C_j in any style.
$c \leftarrow 0$
$t \leftarrow 0$
for all a in A_i **do**
  if $a \in A_j$ **then**
    $c \leftarrow c+1$, $t \leftarrow t+1$, $A_j \leftarrow A_j \setminus a$
  else
    if $a \in A_j^0$ **then** $t \leftarrow t+1$ **end if**
  end if
  $A_i^0 \leftarrow A_i^0 \setminus a$
end for
for all a in A_j **do**
  if $a \in A_j^0$ **then** $t \leftarrow t+1$, $A_i^0 \leftarrow A_i^0 \setminus a$ **end if**
end for
for all a in A_i^0 **do**
  if $a \in A_j^0$ **then** $t \leftarrow t+1$ **end if**
end for
if $t = 0$ **then**
  return *null* {The dataset does not contain authors to compute the similarity between given character styles.}
else
  return c/t
end if

User-Based Similarity In analogy with the style similarity, the user similarity measures the ratio of the number of classes written in the same character style to the total number of common classes provided by two authors.

It helps to determine whether for a given user there are other individuals who have similar writing styles and to suggest the character styles available from those individuals to the given user.

Prediction of Character Style Let $P(S_0|S_1, S_2, ..., S_n)$ be the conditional probability that the character C_0 is written in style S_0 given that the user has provided character styles $S_1, S_2, ..., S_n$. Then for a given symbol, the character style that is suggested to the user at the training phase can be found as

$$\max_{S' \in S} P(S'|S_1, S_2, ..., S_n) \tag{1}$$

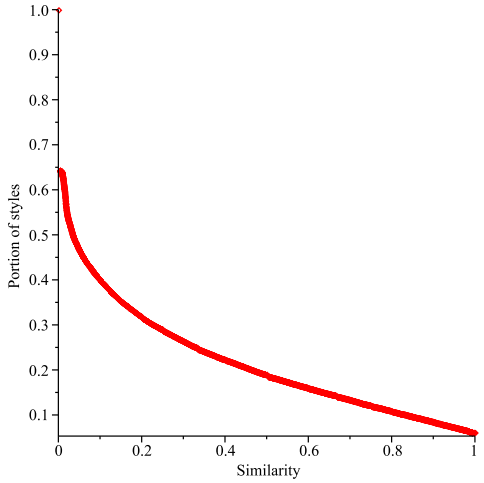

Fig. 3. Portion of pairs of character styles with similarity $\geq$ a given value

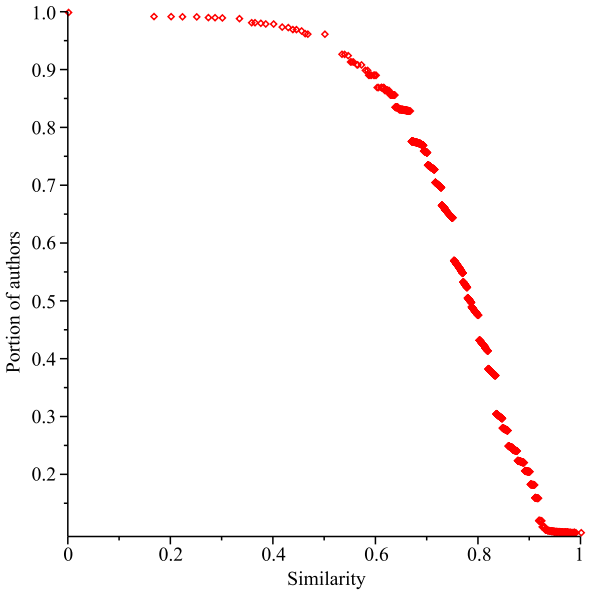

Fig. 4. Portion of pairs of writers with similarity $\geq$ a given value

where S is the set of character styles with which the subject symbol can be written. It can be computed with the chain rule

$$P(\cap_{k=1}^{n} S_k) = \prod_{k=1}^{n} P(S_k | \cap_{j=1}^{k-1} S_j)$$

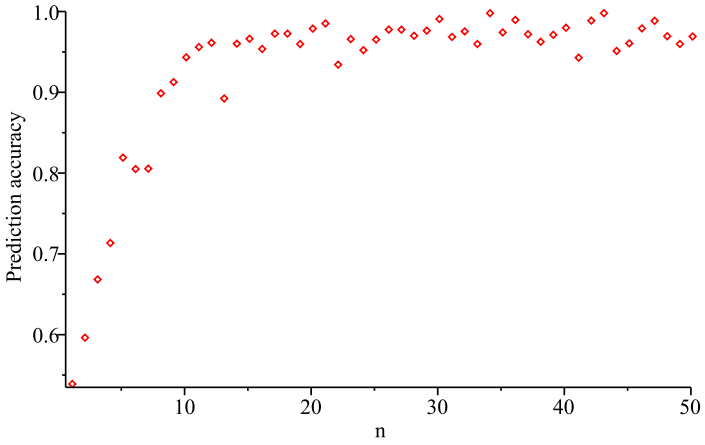

Fig. 5. The character style prediction accuracy

The probability of the user to write n given character styles can be given as

$$P(\cap_{k=1}^{n} S_k)$$

and computed as the ratio of the number of authors who write each of the classes in the corresponding character style to the total number of authors who provided samples for all of the corresponding characters.

4 Experimental Evaluation

In this section we present experimental results. The data set used for testing consisted of 50,703 individual handwritten characters in 242 classes, including Latin and Greek letters as well as mathematical symbols to take into account different forms and styles, as described in [3]. Further, each sample is labeled with its style and the author who provided the sample. There are 369 writers in total.

For the style similarity, we obtained results demonstrated in Figure 3, which shows the portion of pairs of character styles with similarity greater than of equal to a given value. The similarity was found between all combinations of pairs of styles in the collection. The portion is computed as the ratio of the number of pairs of styles with similarity greater than or equal to the given value to the total number of pairs of styles.

Writer similarity is presented in Figure 4. It shows the portion of authors with similarity greater than or equal to a given similarity. The similarity was computed between all combinations of pairs of authors in the dataset. As it was described for the style similarity, the portion is computed as the ratio of the number of pairs of authors with similarity greater than or equal to the given value to the total number of pairs of authors.

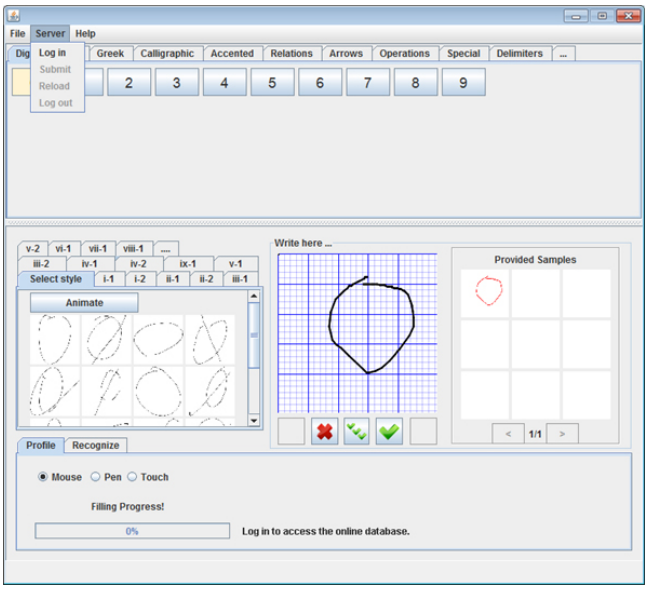

Fig. 6. The training application

For the estimation of the character style prediction accuracy, the experimental runs were organized as follows. For each author, we randomized the list of character styles that the author provided. Then, for each style in the random list, we compute the conditional probability that the corresponding character is written in given style. Figure 5 presents the average prediction accuracy among all writers depending on the number of character styles n available from the author. From the results we can conclude that once an author provided more than 10 styles, we can predict with high accuracy what corresponding character styles the author will be using for other symbols.

5 Use Case: Training a Math Character Recognizer

We now describe an application of the style recommendation algorithm. Consider an application for training a recognizer, developed in our framework for pen-based multi-user online collaboration in mathematical domains [4]. This application, a screenshot of which is in Figure 6, is implemented as an extension of the framework. The extension is designed to collect and organize the training samples in character styles, symbols and catalogs as it is explained in Section 2. This training application is the subject for improvement by asking the user to select the styles suggested by the algorithm, that we present in this paper. Using the idea of style recommendation, the application can be enhanced to suggest styles and corresponding samples to a user, based on the history of styles that the user provided. The UI can be adjusted accordingly. This can speed up the

training of a classifier, because new writers can simply accept the character styles that represent their handwriting and use samples from those styles to train the recognizer.

In concrete terms, our mathematical handwriting database contains 242 classes, and for best results 20 or 30 training samples are required. Although authors may use general, writer-independent recognition, some will want specialized, writer-specific training. With 242 classes, an author who wishes writer-specific recognition would have to give on the order of 5000 to 7000 samples, which is more than most users would be willing to do. Using the recommendation approach described here, a user's style could be detected without having to do full training.

6 Conclusion

We explained the structure of the training dataset, used in our recognition framework. We also briefly described the application for training the classifier. We presented preliminary results of applicability of ideas of recommendation systems to recognition of handwritten mathematical characters. In particular, we performed experiments for estimation of similarity of character styles with respect to writers who provided them, as well as estimation of similarity of writers with respect to their writing styles. Further, we proposed a method for semi-automated training of the classifier that can be used to enhance the described training application. The empirical evaluation demonstrates that about 95% accuracy of prediction of character styles from the writing style of an author can be achieved given 10 character styles from the user.

References

1. Ansari, A., Essegaier, S., Kohli, R.: Internet Recommendation Systems. Journal of Marketing Research 37(3), 363–375 (Aug 2000)
2. Crettez, J.P.: A set of handwriting families: style recognition. In: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. vol. 1, pp. 489–494 vol.1 (1995)
3. Golubitsky, O., Watt, S.M.: Distance-based classification of handwritten symbols. International J. Document Analysis and Recognition 13(2), 133–146 (2010)
4. Hu, R., Mazalov, V., Watt, S.M.: A streaming digital ink framework for multi-party collaboration. In: Proceedings of the 11th international conference on Intelligent Computer Mathematics. pp. 81–95. CICM'12, Springer-Verlag, Berlin, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-31374-5_6
5. Mazalov, V., Watt, S.M.: Writing on clouds. In: Proceedings of the 11th international conference on Intelligent Computer Mathematics. pp. 402–416. CICM'12, Springer-Verlag, Berlin, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-31374-5_27
6. Papagelis, M., Plexousakis, D.: Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents. Eng. Appl. Artif. Intell. 18(7), 781–789 (Oct 2005), http://dx.doi.org/10.1016/j.engappai.2005.06.010