ON-LINE CHARACTER RECOGNITION USING HISTOGRAMS OF FEATURES AND AN ASSOCIATIVE MEMORY

N. Mezghani, A. Mitiche

INRS-EMT
800, de la Gauchetière Ouest, Suite 6900
Montréal QC, H5A 1K6 Canada.

M. Cheriet

École de Technologie Supérieure
1100 rue Notre-Dame Ouest
Montréal QC, H3C 1K3 Canada

ABSTRACT

The purpose of this study is to investigate a new representation of shape and its use in handwritten on-line character recognition. This representation is based on the empirical distribution of features such as tangents, and tangent differences at distant points along the character signal. Recognition is carried out by an associative memory trained using this representation and the Hellinger distance which measures distance between distributions. We report on extensive experiments that show the pertinence of the representation and the superior performance of the scheme.

1. INTRODUCTION

Due to the increasing popularity of hand-held computers, digital notebooks, and advanced cellular phones, automatic on-line recognition of handwritten text has been gaining more interest lately. The traditional methods of human-machine communication, such as keyboards and pointing devices, are often inconvenient to use when the size of the machine is so small as to fit a palm. In this context, handwriting recognition is a very attractive input method.

The most challenging issues in handwriting recognition is the vast variation in personal writing styles. A recognition system should be to some extent insensitive to such variations and still be able to distinguish different but sometimes similar looking characters. Current systems usually have problems handling such variations.

Several methods have been developed for on-line character recognition system [1]. Research has concentrated on classifiers rather than representation, although a good representation is as important as a good classifier.

Various classifiers have been investigated, the most successful being neural networks (TDNN [2], MLP [3]) for reasons such as simplicity of conception, speed of execution, and performance. Hidden Markov models and hybrid HMM/NN have also been used [4, 5, 6, 7]. Finally, nearest neighbor classifiers using elastic matching [8, 9] and weighted elastic matching [10] have been considered in some character recognition applications.

As representation, stroke structures have been the most extensively used [11, 2]. Also topological and geometrical features such as Fourier descriptors have been proposed in [12, 13]. In other studies there were efforts to model pen-tip movement [14, 15] to extract features such as curvilinear and angular velocities.

In this study we propose a new representation of character shape and investigate its use in on-line Arabic character recognition by a Kohonen associative memory. A Kohonen memory is a high performance classifier which requires light training efforts and has attractive properties such as good generalization. As representation we investigate statistics based on features empirical distributions (histograms). The features considered are tangent and tangent differences at regularly spaced points along the character signal.

This paper is organized as follows: section 2 discusses the proposed representation. Section 3 explains the associative Kohonen memory. In section 4 we presents the database and in section 5 we discusses some experimental results.

2. FEATURE REPRESENTATION

Feature extraction and selection are important in achieving high recognition. We investigate a representation based on feature empirical distributions (histograms). The features considered are tangents, and tangent differences at regularly spaced points along the character signal.

Let $\Gamma(s)$ be the arc length parametric representation of the curve of a character, and $N$ equidistant points on $\Gamma(s)$ $(x_0, y_0),(x_1, y_1), ..., (x_k, y_k), ..., (x_{N-1}, y_{N-1})$.

Let $\phi^{(0)}$ be the feature the measurements of which are the tangent angles defined by:

$$\theta_k = \arctan\left(\frac{y_k+1 - y_k}{x_{k+1} - x_k}\right) \quad k \in \{0, 1, ..., N\} \quad (1)$$

For $\alpha \in \mathbb{N}, 1 \leq \alpha \leq N - 1$, let $\phi^{(\alpha)}$ be the feature the measurements of which are the tangent angle differences defined by:
Fig. 1. Tangent angle and tangent angle difference.

\[ \beta_k^{(\alpha)} = \theta_{(k+\alpha)\mod N} - \theta_k \quad k \in \{0, 1, \ldots, N\} \]  

Therefore, for each shape \( \Gamma \), we have a set of features \( \Phi = \{\phi_1^{(\alpha)}, \alpha = 0, 1, 2, \ldots, N-1\} \). The tangent angle \( \phi^{(0)} \) is invariant under translation and scaling and the tangent angle difference is invariant to rotation as well.

Now, we compute statistics for each feature \( \phi^{(\alpha)} \), namely the empirical distribution (histograms) of features. In the continuous case the histogram of \( \phi^{(\alpha)} \) on shape \( \Gamma(s) \) is defined as:

\[ H^{(\alpha)}(\Gamma, z) = \int \delta(z - \phi^{(\alpha)}(s))ds \quad \alpha = 0, 1, \ldots, N-1 \]  

In the above definition, \( z \) is a continuous variable for the feature. For example, \( H^{(0)}(\Gamma, 0) \) is the number of points on \( \Gamma \) that have zero angle. \( \delta \) is the Dirac delta function with unit mass at zero and \( \delta(x) = 0 \) for \( x \neq 0 \). For discretized variables we have:

\[ H^{(\alpha)}(\Gamma, z) = \frac{1}{N} \sum_{j=1}^{N} \delta(z - \phi^{(\alpha)}(s_j)) \]  

In practice, we discretized a histogram \( H^{(\alpha)}(\Gamma, z) \) into \( m \) bins as shown in figure 2. Therefore, for each feature \( \phi^{(\alpha)} \) the histogram is an \( m \)-dimensional vector:

\[ H^{(\alpha)}(\Gamma) = (H_1^{(\alpha)}, H_2^{(\alpha)}, \ldots, H_m^{(\alpha)}) \]

Finally, we compose a vector of representation using histograms of all the features:

\[ \mathcal{H} = (H^{(0)}, H^{(1)}, \ldots, H^{(N-1)}) \]  

3. THE KOHONEN MEMORY

The Kohonen memory (also called Kohonen self organizing feature map) used to represent all points in a source into a lesser number of points in a target space, in such a way that neighboring memory points have neighboring values to preserve topological ordering.

\[ \beta_k^{(\alpha)} = \theta_{(k+\alpha)\mod N} - \theta_k \quad k \in \{0, 1, \ldots, N\} \]  

The memory is organized as a two-dimensional array as shown in figure 3. \( X = (x_1, x_2, \ldots, x_J) \) is the input. Memory node \( j \) stores weight vector \( W_j = (w_{1j}, w_{2j}, \ldots, w_{Jj}) \) determined during training.

The Euclidean distance has been traditionally used in the Kohonen memory algorithm. However, we use the Hellinger distance. The Hellinger distance is a more appropriate measure because it measures distance between distributions.

1. Initialize weights \( W_j \) to small random values, \( j \in [1, J] \).

2. Get new input \( X^n = (x_1^n, \ldots, x_J^n)^T \), and compute the Hellinger distances to all weight vectors \( W_j^n = (w_{1j}^n, \ldots, w_{Jj}^n)^T \) according:

\[ d(X, W_j) = \sum_{i=0}^{J} (\sqrt{x_i^n} - \sqrt{w_{ij}^n})^2 \]
3. Find node $j^*$ with smallest distance.

4. Update weights:

$$w_{ij}^{n+1} = w_{ij}^n + \epsilon_n h_{j^*} \left( x_i^n - w_{ij}^n \right)$$  \hspace{1cm} (7)

$$\epsilon_n = \epsilon_1 \left( \frac{\epsilon_i}{\epsilon_f} \right)^{\frac{n}{\text{max}}}, \quad \sigma_n = \sigma_1 \left( \frac{\sigma_f}{\sigma_i} \right)^{\frac{n}{\text{max}}}$$  \hspace{1cm} (8)

$$h_{j^*} = \exp \left( - \frac{|j - j^*|^2}{2\sigma_n^2} \right)$$  \hspace{1cm} (9)

In training, a set of vectors $X = (x_1, x_2, \ldots, x_J)^T$ is input repeatedly to the $J$ nodes of memory containing each a weight vector $W_j = (w_{1j}, w_{2j}, \ldots, w_{Jj})^T$, initially consisting of random values. Nodes respond to the input vector according to the distance between the input vector and the node's weight vector. The node with weight closest to the input is determined and weights at all nodes are updated. Function $h_{j^*}$ defines the influence of node $j^*$ on node $j$ during update at $j$. It depends on parameter $\sigma$ which decreases with iterations between the value $\sigma_i$ and $\sigma_f$. $\epsilon$ scales weight change and varies with iterations from $\epsilon_1$ to $\epsilon_f$. Parameters $\sigma_i$, $\sigma_f$, $\epsilon_i$ and $\epsilon_f$ must be chosen appropriately to obtain convergence and topological ordering. The result of training is a set of weight vectors $W_j = (w_{1j}, w_{2j}, \ldots, w_{Jj})$, stored at nodes $j = 1, \ldots, J$ to represent the Kohonen map contents. Once the memory is trained, observed vector content is replaced with a sequence of points having the same spatial distance. Therefore, all characters will have the same number of points.

5. EXPERIMENTAL RESULTS

The database is divided into two distinct sets. The training set contained 4896 samples and the testing set 2448 samples (which corresponds to 288 samples of each character for training and 144 samples of each character for testing).

Using all the features $\phi^{(\alpha)}$, $\alpha \in \{0, 1, \ldots, N-1\}$, results in a vector of significantly high dimension (e.g., for $N = 100$ and $m = 10$, the dimension is 1000). Therefore, we use a subset of these features that we have choose experimentally.

We tested these using a memory of 400 nodes (this is a sufficient size, also determined experimentally) and a number of iterations of 80. Table 1 shows the recognition rates for $\alpha$'s multiples of 10. The tendency is for the rate, as a function of $\alpha$, to grow to maximum and then decrease.

<table>
<thead>
<tr>
<th>Index feature $\alpha$</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate $\tau$</td>
<td>68.99</td>
<td>66.54</td>
<td>70.38</td>
<td>75.32</td>
<td>77.04</td>
</tr>
<tr>
<td>Index feature $\alpha$</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>Recognition rate $\tau$</td>
<td>74.95</td>
<td>69.32</td>
<td>70.38</td>
<td>75.32</td>
<td>77.04</td>
</tr>
</tbody>
</table>

Table 1. Recognition rates vs. index features (rates obtained for each histogram taken individually)

Following these first experiments, we retained the features for $\alpha = 0, 10, 20, 30, 40$, to compose the vector of representation which, in this case, is of dimension 50. With this vector of representation we obtained a superior recognition rate of 94.56%. The size of the memory and the number of iteration have also been selected experimentally. The retained number of nodes is 400 and the number of iterations 80 (Fig. 5).

As one can notice, the recognition rate obtained by the resulting vector of representation is far superior than any histogram taken individually.
We also measured the various execution times. Table 2 shows that the recognition time is relatively short. This is very important for an on-line recognition system. The training time is long, but is not a disadvantage because training is done only one time.

<table>
<thead>
<tr>
<th>steps</th>
<th>Training</th>
<th>Labeling</th>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Execution time</td>
<td>1h 5mn</td>
<td>8.47s</td>
<td>1.05s</td>
</tr>
<tr>
<td>per character</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Execution time vs. steps in Kohonen map

6. CONCLUSION

The aim of this paper was to develop a new representation of shape and to use it in handwritten on-line Arabic character recognition. We investigated statistics of feature based on histograms of tangent angles and tangent angle differences. Recognition was carried out by an associative memory. Experimental results show the high performance of the proposed scheme and the pertinence of the representation.

7. REFERENCES


