Fuzzy Point of View Combination for Contextual Shape Recognition: Application to On-line Graphic Gesture Recognition

François Bouteruche  Éric Anquetil
IRISA\INSA de Rennes
Campus universitaire de Beaulieu
35042 Rennes Cedex, France
{francois.bouteruche, eric.anquetil}@irisa.fr

Abstract

In this paper, we focus on the explicit use of the spatial context of strokes in the recognition process of graphic gestures. We dispose of two sources of knowledge on the samples: their shape which is a classical one and their spatial context. The proposed method is based on three different points of view to exploit optimally these two sources of knowledge to perform the recognition. The first point of view uses the spatial context to filter the possible classes and then uses the shape to discriminate the remaining classes. The second one reverses the roles of the sources of knowledge. The third one uses them jointly. The underlying idea is that each point of view suits to one of the three case of source reliability. The challenge is to automatically compose the points of view and to combine them to build a system performing contextual shape recognition without any prior information on the targeted domain.

1. Introduction

On-line handwriting recognition systems now reach performances and reliability that allow using them in real-word pen-based software. However they are only able to recognize isolated entities (words, characters…), so they don’t completely fit the needs. For instance, in a handwriting note-taking application with sketch recognition capabilities, it is impossible to discriminate strokes representing circles, letters ‘o’ and numbers ‘0’ without any information on their context to select the appropriate dedicated recognizer.

In this paper, we focus on the explicit exploitation of the spatial context in the recognition process. Actually, few works address this problem. In [3, 6, 8], it is viewed as information on the structure of the symbol. They use it either jointly with shape information in the recognition system or after the symbol recognition to resolve ambiguities. Most of the time in this case, an expert defines the rules of the spatial context use for each targeted applications.

Our goal is to incorporate the spatial context in the learning process to build recognizers that automatically take advantage of it. In [2], we proposed a first method that determines the different spatial contexts and learns the associated shape recognizers from training data. During the recognition, the spatial context of the stroke is compared to the different spatial context models and the shape recognizer associated to the most relevant one is called. This method is similar to a system using expert’s rules to define the contexts, but the expert’s rules are replaced by an automatic learning process.

This hierarchical approach is based on the idea that only a subset of all possible symbols can be written in each specific spatial context. For example, if a stroke is written above a previously recognized Latin character, it is more likely a diacritic than another character. So, the recognition system can focus on diacritic recognition. Despite the intuitive aspect of this approach, one of its major problems is the a priori choice of the knowledge hierarchical order.

Considering this, we propose a new approach using three points of view based on three different compositions of two sources of knowledge: the spatial context and the shape. The first one uses the spatial context to focus the system on groups of symbols sharing similar contexts and based its final decision on the shape. On the contrary, the second point of view uses the shape to focus the system on groups of symbols sharing similar shapes and based its final decision on the spatial context. The third point of view uses jointly both the sources of knowledge to take its decision. The global system relies on these three points of view that we combine to optimize the global decision. The general framework of the approach relies on the fuzzy sets theory.

The main challenge of this work is to automatically generate the points of view and combine them in order to optimize the global performance of the system. The aim is
to propose a method allowing extracting and combining knowledge on the spatial context and the shape to perform contextual shape recognition without any prior information on the targeted domain. Moreover the system must use few memories since its targeted platform is the small-size devices such as smartphones or Personal Digital Assistants (PDA). To validate the method, we apply it to on-line graphic gesture recognition in an input method for PDA. These gestures allow editing the previously inputted characters.

In section 2, we explain how we compose the points of view from both sources of knowledge. Then, section 3 presents the contribution of using three points of view and how we combine them. Finally, we present the first experimental results in section 4 and conclude in section 5.

2. Composition of points of view

2.1. Principle

To propose a system able to automatically extract and combine the spatial context and the shape to perform recognition, we need a supervised data set to learn its models. Consequently in this work, the class of each training sample is known. Moreover, we consider that each training sample can be described by two kinds of information. The first one is its shape defined by a set of features describing its morphological properties (its width, its height . . . ). The second one is its spatial context defined by a set of features describing its position relatively to a reference. We assume here that the reference can be previously determined.

As each sample is described by two feature sets into two distinct feature spaces, we have two distinct sources of knowledge on each sample. We use them to extract the intrinsic properties of each class in each feature space independently of the other classes. Once the intrinsic properties of each class are extracted, we dispose of two distinct sources of knowledge on each class.

From these sources of knowledge, two different kinds of point of view can be learned. The first one is based on the hierarchical use of the knowledge. The second one is based on the joint use of the knowledge.

The hierarchical points of view rely on the idea that given a specific area of one of the feature spaces only a subset of all classes are interested in. Hence there is no need to consider all the classes in the other feature space: the system can focus on the subset of classes. If we consider a hierarchy that first uses the spatial context and then uses the shape, it means that in a specific spatial context only a subset of all possible symbols with distinct shapes can be drawn.

The joint point of view relies on the idea that when the knowledge on the spatial context and the shape are not very reliable, the class of the sample can be determined using jointly these two sources of knowledge. It means that although the spatial context and the shape of a sample can not be identified accurately, the conjunction of these two pieces of information allow to optimizing the decision.

Section 2.2 presents how we extract the knowledge on the spatial contexts and the shapes of the classes. Section 2.3 and 2.4 explains how to combine this knowledge to compose a hierarchical point of view and a joint one. Section 2.5 explains how to exploit them for recognition.

2.2. Knowledge extraction

To extract the intrinsic properties of the classes, we build two data sets for each of them. The first one contains the examples described by their spatial context features and the second one the same examples described by their shape features. Then we apply on these sets a clustering algorithm.

Several clustering algorithms that find relevant prototypes exist. Here, we use the Possibilistic C-Means fuzzy clustering algorithm [5]. It describes the clusters by independent fuzzy prototypes \( P_i^j \) defined by their centers \( \vec{c}_j^i \) and their membership functions \( \mu_i^j \). The used functions are hyper-ellipsoidal radial basis functions with \( \vec{c}_j^i \) as center. Their shapes are given by the covariance matrix \( Q_j^i \) using the Mahalanobis distance:

\[
\mu_i^j(\vec{X}) = \frac{1}{1 + \frac{d_{Q_j^i}(\vec{X}, \vec{c}_j^i)}{2}} \tag{1}
\]

The obtained fuzzy prototypes on the data sets model our knowledge on the spatial contexts and the shapes of the classes. A class \( i \) is characterized by two sets of fuzzy prototypes \( P_{c_i}^j \) and \( P_{s_i}^l \) with \( 1 \leq i \leq C, 1 \leq j \leq K_1 \) and \( 1 \leq l \leq K_2 \). \( C \) is the number of classes, \( K_1 \) and \( K_2 \) the numbers of clusters used to model a class in each feature spaces. The fuzzy prototypes \( P_{c_i}^j \) model the interest areas of the class \( i \) in the context feature space, whereas the fuzzy prototypes \( P_{s_i}^l \) model its interest areas in the shape one. The membership functions \( \mu_{c_i}^j \) of \( P_{c_i}^j \) and \( \mu_{s_i}^l \) of \( P_{s_i}^l \) allow computing the membership value of a sample to the corresponding area of interest. This value represents the typicality of the sample referring to this area. Hence, \( \mu_{c_i}^j \) evaluates how typical of the specific context \( j \) of the class \( i \) a sample is. \( \mu_{s_i}^l \) evaluates how typical of the specific shape \( l \) of the class \( i \) a sample is.

2.3. Hierarchical point of view

The hierarchical knowledge combination consists in determining which classes are activated in each interest areas of the feature space chosen as first. For example, if the context feature space is this one, we must learn which classes can be realized in each specific contexts described by the fuzzy prototypes \( P_{c_i}^j \) and more precisely which specific shape of these classes. So, we must learn the associations between each \( P_{c_i}^j \) and all \( P_{s_k}^l \) describing the shape
of the classes $k$ that can be realized in this specific context. To do so, we associate to each fuzzy prototype $Pc_i^c$, the set $A(Pc_i^c)$ of fuzzy prototypes $Ps_k^l$ such as:

$$A(Pc_i^c) = \left\{ Ps_k^l \mid \exists X = (\hat{X}c, \hat{X}s) \in TrainingSet \wedge \mu_{c_i}^l(\hat{X}_c) = \max_{1 \leq n \leq K_1} \mu_{c_m}^n(\hat{X}_c) \wedge \mu_{s_k}^l(\hat{X}_s) = \max_{1 \leq n \leq K_2} \mu_{s_m}^n(\hat{X}_s) \right\} \quad (2)$$

It is important to note that a fuzzy prototype $Ps_k^l$ describing the shape of a class can be associated to several fuzzy prototypes $Pc_i^c$. It means that this class can be drawn in several spatial contexts. Moreover, all the fuzzy prototypes $Ps_k^l$ describing the shape of a class are not necessarily associated to the same fuzzy prototype $Pc_i^c$ describing a spatial context. So in a specific spatial context, it is possible that only a specific shape of the class can be drawn.

### 2.4. Joint point of view

The joint knowledge use combination consists in computing the membership value of a sample to the context and the shape of each class $i$ using the $\max$ t-conorm:

$$\mu_{c_i}(\hat{X}_c) = \max_{1 \leq j \leq K_1} \mu_{c_j}(\hat{X}_c) \quad (3)$$

and then to compute the typicality of the sample to each class $i$ thanks to a fuzzy conjunction using the $\min$ t-norm:

$$\mu_i(\hat{X}) = \min(\mu_{c_i}(\hat{X}_c), \mu_{s_i}(\hat{X}_s)) \quad (5)$$

### 2.5. Use of a point of view for recognition

The recognition process of a hierarchical point of view composes of two steps. In the case of a point of view using first the spatial context, the system computes the fuzzy prototypes $Pc_m^c$ such as:

$$\mu_{c_m}^n(\hat{X}_c) = \max_{1 \leq n \leq K_1} \mu_{c_m}^n(\hat{X}_c) \quad (6)$$

Then it determines the fuzzy prototypes $Ps_p^l$ such as:

$$\mu_{s_p}^l(\hat{X}_s) = \max_{1 \leq l \leq K_2} \mu_{s_p}^l(\hat{X}_s) \quad (7)$$

The decision of the system is the class $p$ described by $Ps_p^l$.

In the case of a joint point of view, the system computes the typicality $\mu_i(\hat{X})$ of the sample to each class and then determines the class $p$ such as:

$$\mu_p(\hat{X}) = \max_{1 \leq i \leq C} \mu_i(\hat{X}) \quad (8)$$

### 3. Using and combining the three points of view

#### 3.1. Why using the three points of view ?

In most cases the knowledge on the spatial contexts of the samples and the knowledge on their shapes are reliable, but for some samples they are not.

In the case where one of them is less reliable in terms of discrimination, it is wise to base the final discriminatory decision on the most reliable one. Yet the less reliable source can be used robustly to select a subset of all possible classes. This allow resolving the easiest overlap problems in the feature space representing the most reliable source of knowledge and to focus the system on the remaining overlap problems. In the case where both the sources are unreliable, it is wise to use them jointly to take the less bad decision.

Considering this, it appears that a system using only one point of view is suitable only if at least one of the sources is totally reliable, but this is an ideal case. So a wise approach is to use a combination of the three points of view to build a system that suits all the cases.

#### 3.2. Combining the three points of view

The best point of view to recognize a sample can not be chosen a priori. To combine the three points of view, we must evaluate the reliability of the sources of knowledge relatively to the samples. This evaluation is possible during the learning step since the training sample labels are known.

The solution is to start the recognition process using one of the hierarchical points of view since they state that at least one of the sources is reliable. It must evaluate the confidence in its decision and if it is too low, let the second hierarchical point of view classify the sample. If this one evaluate that the confidence in its decision is too low too, it means both the sources are unreliable for this sample. Then, it let the joint point of view take the final decision.

Reliability evaluators [4, 7] are used to measure the confidence in the decision of the two hierarchical point of view. They compute the relative differences between the two most activated fuzzy prototypes of the second feature spaces describing two distinct classes. Reject thresholds on these reliability evaluators are learned to determine if the confidence is too low [7]. They are optimized to maximize the accuracy of the points of view, i.e. to reject the maximum of errors.

### 4. First experimental results

#### 4.1. Performance evaluation

We have tested this new approach on an extended version of the benchmark used in [2]. It consists in the recognition of 18 on-line graphic gestures. 16 of them correspond to an edition of their reference character and 2 of them correspond to the addition of a stroke to the character. The examples have been written on PDA. The training and test data sets contain 4243 examples of 8 writers and 2393 examples of 6 writers respectively. None of the writers is common to both data sets. 4 features describe them in the spatial context feature space and 9 in the shape feature space.
First we tested the need of the spatial context to recognize the 18 graphic gestures. We learned a Radial Basis Function Network (RBFN) [1] using only the shape features. It obtained a 52.90% recognition rate (RBFN 1 in table 1). Then we trained a RBFN (RBFN 2) using both the spatial context features and the shape features. It obtained a 92.73% recognition rate which shows that proper performance can not be reach without the spatial context. It will be our reference to evaluate the proposed method. Finally, we learned our 3 composed point of view system (3CPV). It obtained a 93.56% recognition rate and so reduces the error rate by 11.42% compared to the RBFN 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>RBFN 1</th>
<th>RBFN 2</th>
<th>3CPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition</td>
<td>52.90%</td>
<td>92.73%</td>
<td>93.56%</td>
</tr>
<tr>
<td>Sample number</td>
<td>2393</td>
<td>499</td>
<td>108</td>
</tr>
<tr>
<td>RBFN 2 rate</td>
<td>92.73%</td>
<td>80.56%</td>
<td>51.85%</td>
</tr>
<tr>
<td>Performance</td>
<td>80.07%</td>
<td>83.77%</td>
<td>45.37%</td>
</tr>
<tr>
<td>Rejection rate</td>
<td>20.85%</td>
<td>21.64%</td>
<td>N/A</td>
</tr>
<tr>
<td>Recognition</td>
<td>75.93%</td>
<td>74.75%</td>
<td>N/A</td>
</tr>
<tr>
<td>Accuracy rate</td>
<td>95.93%</td>
<td>95.40%</td>
<td>N/A</td>
</tr>
</tbody>
</table>

4.2. Performance analysis

The first point of view of our system is a hierarchical one (section 2.3) using first the spatial context and then the shape (PV 1 in table 2). It runs on 2393 test samples. Its performance rate on these samples (without reject) is 80.07% whereas the RBFN 2 achieves a 92.73% rate. Yet, thanks to its rejection capabilities, it achieves a 95.93% accuracy rate with a 20.85% rejection rate. It means that it reaches a 95.93% recognition rate on the 1894 samples it classifies and let the 499 other samples to the second point of view.

This second point of view is a hierarchical one using first the shape and then the spatial context (PV 2). It runs on the 499 samples rejected by PV 1. Its performance rate is 83.77% whereas the RBFN 2 rate is 80.56%. It achieves a 95.40% accuracy rate with a 21.64% rejection rate.

The third point of view is the joint one (section 2.4, PV 3). It runs on the 108 samples rejected by PV 2 without rejection capabilities. Its performance rate on these samples is 45.37% whereas the RBFN 2 achieves a 51.85% rate.

5. Conclusion

In this paper, we propose a new approach that incorporates the spatial context of the strokes in the learning process of recognizers. This method is based on the composition of three points of view using two sources of knowledge in three different ways. We present how to combine these points of view to build a system taking the best of them.

The results show that this new method allow better performances than a classical method such as the RBFN thanks to a combination based on the accuracy of the point of view.

Our future works will focus on testing the method on more complex recognition problems, i.e. with more classes and more possible contexts (music score recognition...). Another directions will be the use of weight layers to refine the point of view decision, the study of the rejection strategies and in-depth analysis of combination methods.

References