Facial Expression Classification Using Combined Neural Networks

Rafael V. Santos, Marley M.B.R. Vellasco, Raul Q. Feitosa, Ricardo Tanscheit
DEE/PUC-Rio, Marquês de São Vicente 225, Rio de Janeiro – RJ - Brazil
marley@ele.puc-rio.br

Abstract

Studies in the area of Pattern Recognition have indicated that a classification model performs differently from class to class. This observation leads to combining the individual results of different classifiers to derive a consensus decision. This work investigates the combination of classifiers in a facial expression recognition system. A classifier ensemble is then built by integrating the results of several Back-propagation neural networks; Fuzzy Integrals are used as the combining strategy. Experiments carried out to evaluate the system have shown that the combination may considerably improve the classification performance of the individual classifiers.

1 Introduction

The interest on systems for automatic recognition of facial expressions has increased recently. Such systems are clearly relevant in studies of human behavior, since facial expressions are a manifestation of human emotions. Facial expressions also play an important role in the non-verbal communication between human beings and indeed overcome that of actual words [1]. This has awakened the interest in the development of more effective techniques for computer-human interaction.

Automatic systems for recognition of facial expressions must deal with three basic problems: detection of the human face in a generic image, extraction of relevant attributes from the facial image and the classification itself.

Locating a face in a generic image still constitutes a challenge. Once detected, the image region containing the face is extracted and geometrically normalized, usually maintaining a constant inter-ocular distance. References to detection methods using neural networks and statistical approaches can be found in [2] and [3]. All experiments presented in the next sections use well-framed face images as input.

The second problem is concerned with the selection of a set of attributes that can represent appropriately the emotions expressed on the images. Among the proposed approaches for the selection of attributes [4] the Principal Component Analysis algorithm (PCA) has been widely used [5].

With respect to the third problem, Neural Networks have been successfully used as classifiers on face recognition systems [6]-[8]. In this paper a Neural Network
Ensemble [9],[10] is used for the recognition task. Each of the ensemble’s components is a Back-propagation network, since this is the most widely used algorithm. Fuzzy Integrals [11]-[13] are used as the combining technique. Unlike [6], the system proposed here utilizes well-framed, static images, obtained through a semi-automatic method. Instead of geometrical attributes, the principal components analysis has been applied to generate the vector of relevant attributes.

The remaining of the text is organized as follows. Section 2 describes the whole system, presenting brief descriptions of the designed ensemble’s components and of the combining schemes based on Fuzzy Integrals. Section 3 describes the experiments that have been performed. The results are then shown in Section 4, which is followed by the conclusions in Section 5.

2 Methodology

2.1 System’s General Architecture

The automatic system proposed for the recognition of facial expressions is composed of three stages: detection, extraction of attributes and classification (Figure 1). The second stage, Extraction of Attributes, is performed through the PCA algorithm by exploring the concept of eigenfaces [14].

![Figure 1: Facial Expression Recognition System](image1)

The third stage, classification, is itself divided in two others: individual classification and data fusion (Figure 2). Each individual classifier – a Local Expert Neural Network (section 2.2) - gives a self-sufficient classification result, which is then combined to the others by Fuzzy Integrals (section 2.3). This is the so-called Classifier Ensemble, in the very sense proposed by A. Sharkey in [9].

![Figure 2: Classification Stage](image2)

2.2 Local Expert Neural Networks

The usual approach for Neural Network-based pattern recognition applications consists of training a single network so that it achieves the lowest classification error over all classes. In this work we use a different approach, related to the creation of a Network Ensemble [9],[15]. In order to achieve good classification performance, it is well known that each combined component should be (1)
individually accurate and (2) make their errors on different parts of the input space. Therefore in this work we search for the most competent network for each class, resulting in a set of Local Expert Networks.

Local competence measures per class are defined as [16]:

\[
g_i(x_i) = \frac{C_{ik}}{C_{ik} + \sum_{j \neq k} C_{ij} + \sum_{j \neq k} C_{jk}}
\]

with \( k = 1, \ldots, n \) being the number of classes. In Eq.1 \( C_{ik} \) is the number of patterns in the training set assigned by the classifier \( x_i \) to the class \( t_k \). Therefore \( \sum_{j \neq k} C_{ij} \) is the number of training patterns belonging to the class \( t_k \) and assigned by \( x_i \) to a different class. Similarly \( \sum_{j \neq k} C_{jk} \) is the number of training patterns not belonging to the class \( t_k \) and assigned by \( x_i \) to \( t_k \).

A local expert network for a class \( t_k \) can be built by replicating (by an integer factor \( \gamma \) greater than one) the patterns from \( t_k \) in the chosen training set. In other words, if there are originally \( n \) patterns belonging to \( t_k \) in the training set, there will be \( \gamma n \) after the replication, which will ideally improve the learning process for that class. As a result, a regular Back-propagation training will not only produce the expert but also move towards an overall accurate network.

2.3 Applying Fuzzy Integrals to Classifiers combination

2.3.1 Fuzzy Measures

A fuzzy measure is defined by a function that assigns a value in the [0,1] interval to each crisp set of the universal set [17]. In the context of classifier combination, a fuzzy measure can express the level of competence of a classifier in assigning a pattern to a class. It must be noted that this is different from the concept of membership grade. In the latter case a value is assigned by a classifier to a pattern, expressing its degree of membership to a particular class. The fuzzy measure, on the other hand, denotes the level of trust on this classifier when evaluating the membership degree for a given class.

Formally, a fuzzy measure is a function \( g_{A \subseteq \Omega} : X \rightarrow [0,1] \), where \( \Omega \) is the universal set comprising all crisp sets of a specific variable \( x \).

A fuzzy measure is similar to a probability measure, except that it does not follow the addition rule, that is: if \( g \) is a fuzzy measure defined over a set \( \Omega \) and \( A, B \subseteq \Omega \) so that \( A \cap B = \emptyset \), the equation \( g_{A \cup B}(x_i \cup x_j) = g_{A}(x_i) + g_{B}(x_j) \) does not apply.

2.3.2 Fuzzy Integrals

By using the concepts of fuzzy measures [17], a fuzzy integral has been defined [11] as a non-linear operation defined over measurable sets.

Let \( A \) be an object (pattern) to be classified. Let \( T = \{ t_1, t_2, \ldots, t_n \} \) be the set of possible classes to be chosen and \( X = \{ x_1, x_2, \ldots, x_m \} \) the set of available classifiers.

Fuzzy measures \( g_k(x_i) \), denoting the competence of classifier \( x_i \) in the recognition of patterns belonging to class \( t_k \), must be set to each classifier to be combined. These densities may be set by experts or by training sets analysis. In this paper \( g_k(x_i) \) is the hit ratio at the training phase for classifier \( x_i \) with respect to class \( t_k \).
Let $h_k: X \rightarrow [0,1]$ be a function which expresses how well the pattern fits into the class $t_k$ according to the classifier $x_i \in X$. If the cardinality of $X$ is $m$, then $X$ is arranged as $\{x_1, x_2, \ldots, x_m\}$ so that $h_k(x_1) \geq h_k(x_2) \geq \ldots \geq h_k(x_m) \geq 0$.

An ascending sequence of classifiers $Y=\{y_1, y_2, \ldots, y_n\}$ will then be created, so that $y_1 = x_1$ and $y_i = y_{i-1} \cup x_i$, for $1 < i \leq n$, whereby the symbol $y_{i-1} \cup x_i$ denotes the classifier resulting from the combination of classifier $y_{i-1}$ with classifier $x_i$.

Since fuzzy measures do not follow the addition rule, Sugeno’s proposal is used to compute the fuzzy measures for the new sequence of classifiers, as shown in Eq.2:

$$g_k(y_i) = g_k(y_{i-1} \cup x_i) = g_k(y_{i-1}) + g_k(x_i) + \lambda g_k(y_{i-1}) g_k(x_i), \quad (2)$$

where $\lambda > 0$. The value of $\lambda$ is always taken from the boundary condition $g(y_m) = 1$, which means that the fuzzy measure of the classifier resulting from the combination of all original classifiers will be equal to 1. By using Eq. 2 recursively to describe that condition, $\lambda$ is determined by solving an $n-1$ degree equation:

$$\prod_{i=1}^{n}[1 + \lambda g_k(x_i)] = 1 + \lambda, \lambda \neq 0 \quad (3)$$

Sugeno has proved that there is always a unique (non-zero) $\lambda \epsilon (-1, \infty)$ satisfying Eq.3. The fuzzy integral $(e_k)$ of the function $h_k$ over $Y$ with respect to $g_k$ is given by:

$$e_k = \sqrt[n]{\frac{1}{m} \max_{i=1}^{m} [h_k(x_i) \cdot g_k(y_i)]} \quad (4)$$

This expression is computed in two steps:

1. Obtain the product (or t-norm) between $h_k(x_i)$ and $g_k(y_i)$, for $1 \leq i \leq m$ and
2. Determine the maximum (or t-conorm) of the resulting sequence from phase 1.

There are several interpretations for fuzzy integrals; here it is useful to see them as a method for obtaining the maximum grade of agreement between competence $g_k(y_i)$ and confidence $h_k(x_i)$. According to this procedure, a pattern will be assigned to the class having the highest value returned by the fuzzy integral.

As will be noticed later, the aforementioned competence measures defined for each local expert network will be used in this work as the fuzzy measure per class associated to each classifier.

3 Experiments

3.1 Image database

The database used in this work has been extracted from ATR Human Information Processing Research Labs. It consists of 160 grayscale images of 9 Japanese women. The database contains pictures with the six basic facial expressions, as defined in the Facial Action Coding System (FACS) [18]: disgust, fear, anger, sadness, surprise and happiness. There are three or four images for each expression for each person.

The figure below illustrates the six types of expressions used in this work.
A large portion of the images of this database carries no relevant information for the classification procedure. Thus the image region covered by the face was extracted from the complete image, as shown in Figure 4. This could be easily implemented given that all images were well centered. After extraction, the faces were resized to 64x64 pixels, in order to reduce computational load.

3.2 Implementations

Two subsets have been built from the original database: one for training, another for validation. Those subsets have been composed of exactly one picture per class per person, randomly, achieving a total of 54 patterns. The remaining 106 observations formed the training set.

The networks used here have a single hidden layer. Six log-sigmoid were designed to represent the expected class of a given pattern; that is, an output vector with values [0 0 1 0 0 0] would indicate that the pattern belongs to the third class. The number of neurons in the hidden layer was set to 5, and the number of inputs (or principal components extracted from the input data), was set to 15. These numbers have been chosen by following the reasoning: the number of training patterns should be equal or greater the number of network weights \((w)\) to be adjusted. Considering a fully-connected network, \(w = h(i + o)\), where \(h\) is the number of hidden layer neurons, \(i\) the number of inputs and \(o\) the number of outputs. In our case: \(106 \geq h(i + 6)\), or \(106 \geq 5(15+6)\), which certainly holds. Finally, the training and validation sets were transformed to standard scores (z scores) [19].

The Back-propagation parameters were a momentum constant of 0.9 and a learning rate of 0.2. All networks were trained for 1,000 epochs. Despite this fixed number, it should be mentioned that the best networks chosen after each experiment were
those with the lowest *Mean of Squared Errors* (MSE) at the validation set. In other words, while trying to reduce the MSE at the training set (the usual procedure), the performance was also being checked at the validation one, so that the *best* network would be chosen. As our main objective is to analyze the overall classification performance of Ensembles against Single Classifiers, no great attention was paid to any individual performance issue, in both approaches. Future works, however, must consider a complete cross-validation scheme to improve generalization confidence.

4 Results

The first executed test consisted of training a non-expert network classifier for the facial expressions problem. This was the same as to apply $\gamma = 1$ to the training set (section 2.2). After 1000 epochs the following network has been selected (expressed at Table 1 in terms of hits (*recognition rates*) per class over the validation set):

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
<th>Average Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>88.9</td>
<td>77.8</td>
<td>88.9</td>
<td>88.9</td>
<td>77.78</td>
<td>100</td>
<td>87.04</td>
</tr>
</tbody>
</table>

Once this was established as reference for comparison, the next step consisted of training the six expert networks to be combined (NN1, NN2, ..., NN6), each one being expert (in comparison to the others) in one of the existing classes. Table 2 shows the competence measured at the training phase for the individual classifiers.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN1</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>0.95</td>
<td>0.889</td>
<td>0.85</td>
</tr>
<tr>
<td>NN2</td>
<td>0.84</td>
<td>1.00</td>
<td>0.95</td>
<td>0.78</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>NN3</td>
<td>0</td>
<td>0.76</td>
<td>1.00</td>
<td>0.82</td>
<td>0.72</td>
<td>0.81</td>
</tr>
<tr>
<td>NN4</td>
<td>1</td>
<td>0.82</td>
<td>1.00</td>
<td>1.00</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>NN5</td>
<td>0</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.99</td>
<td>0.60</td>
</tr>
<tr>
<td>NN6</td>
<td>0.76</td>
<td>0.78</td>
<td>0.86</td>
<td>0.61</td>
<td>0.80</td>
<td>0.99</td>
</tr>
</tbody>
</table>

The bold values in Table 2 indicate the networks expertise. The last step was to perform the information fusion from the individual classifiers. This was accomplished by using the concepts of section 2.3 and led to the recognition rates showed in the last row (MIX) of Table 3. The first six rows report hits from the individual networks, as a matter of comparison.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN1</td>
<td>100</td>
<td>77</td>
<td>89</td>
<td>100</td>
<td>67</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td>NN2</td>
<td>78</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>56</td>
<td>100</td>
<td>83</td>
</tr>
<tr>
<td>NN3</td>
<td>0</td>
<td>89</td>
<td>78</td>
<td>100</td>
<td>78</td>
<td>100</td>
<td>74</td>
</tr>
<tr>
<td>NN4</td>
<td>89</td>
<td>78</td>
<td>89</td>
<td>100</td>
<td>67</td>
<td>100</td>
<td>87</td>
</tr>
<tr>
<td>NN5</td>
<td>0</td>
<td>78</td>
<td>78</td>
<td>100</td>
<td>100</td>
<td>89</td>
<td>74</td>
</tr>
<tr>
<td>NN6</td>
<td>67</td>
<td>67</td>
<td>78</td>
<td>56</td>
<td>67</td>
<td>100</td>
<td>72</td>
</tr>
<tr>
<td>MIX</td>
<td>89</td>
<td>89</td>
<td>100</td>
<td>89</td>
<td>89</td>
<td>100</td>
<td>93</td>
</tr>
</tbody>
</table>

By comparing Tables 1 and 3, it can be seen that the classification performance of the assemble approach provided better performance (or at least the same performance) in all classes, proving that the combination of expert networks and the fuzzy integral fusion process is a promising ensemble methodology.
5 Conclusions

The potential of combining classifier to improve classification accuracy of facial expressions images has been investigated. A classification system was proposed, which combines the results of local experts neural networks. Fuzzy integrals were used as combination strategy.

The system was evaluated on a well-known expressions database. In the experiments for performance evaluation the combination attained an average performance considerably higher than that of individual classifiers. The experiments have also shown that the combination tends to equalize the performance among all classes, while improving the overall recognition rate.

Results encourage a further investigation of combined classifiers for this kind of application, in order to obtain a deeper understanding on combination strategies and in what circumstances they may be employed.

Acknowledgments

We are thankful to ATR Human Information Processing Research Labs for making available the face database for our experiments.

References