An Evolving Neural Network Model for Person Verification Combining Speech and Image

Akbar Ghobakhlou, David Zhang, and Nikola Kasabov

KEDRI, Auckland University of Technology, Private Bag 92006, Auckland, New Zealand
{akbar,dzhang,nkasabov}@aut.ac.nz

Abstract. This paper introduces a method based on Evolving Connectionist Systems (ECOS) for person verification tasks. The method allows for the development of models of persons and their on-going adjustment based on new speech and face images. Some experimental person verification models based on speech and face image features are developed based on this method where speech and face image information are integrated at a feature level to model each person. It is shown that the integration of speech and image features improves significantly the accuracy of the person verification model when compared with the use of only image or speech data.

1 Introduction

Biometric verification can be defined as a process of uniquely identifying a person by evaluating one or more distinguishing biological traits. Unique identifiers include fingerprints, hand geometry, retina, iris patterns, face image and voice. There are many biometric features that distinguish individuals from each other, thus many different sensing modalities have been developed [5]. These identifiers may be used individually, as exemplified by the iris scan system deployed in the banking sector and currently being tested for airport security [6].

Over the past few years, interest has been growing in the use of multiple modalities to solve automatic person identification problems. The motivation for using multiple modalities is multi-fold. In the first instance different modalities measure complementary information and by this virtue multimodal systems can achieve better performance than single modalities. Single feature may fail to be exact enough for identification of individuals.

In this paper we propose person verification module based on Evolving Connectionist Systems (ECoS) [1]. Person verification models developed based on speech face image and integrated features. Each person is modelled by placing nodes during the training process. From previous work [2], it was shown that ECoS could be used to create adaptive speech recognition systems. ECoS use a local learning algorithm where each neuron in the evolving layer of the network represents data in a small region from the problem space.

The following sections describe the method used and the experimental system built to demonstrate the method. First, the pre-processing and feature extraction methods are described and then the classification ECoS principles are presented and illustrated on a simple case problem.

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 381–386, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Speech and Face Image Signal Processing

2.1 Speech Signal Sampling and Processing

In the speaker verification model, a text-dependent module was built. The speech data was captured using close-mouth microphone. The speech was sampled at 22.05 kHz and quantized to a 16 bit signed number. In order to extract Mel Frequency Cepstrum Coefficients (MFCC) as acoustic features, spectral analysis of the speech signal was performed over 20ms with Hamming window and 50% overlap. Discrete Cosine Transformation (DCT) was applied on the MFCC of the whole word to obtain input feature vectors [2].

2.2 Face Image Processing

In the face verification model, the images were captured using a web-cam with a resolution of 320×240. Once a new image was captured, features were extracted using the composite profile technique. The composite profile features are composed of the average value of the columns in the image followed by the average value of rows in the image. It is a relevant feature to characterize symmetric and circular patterns, or patterns isolated in a uniform background. This feature can be useful to verify the alignment of objects. In order to reduce the number of features, the interpolation technique was applied to the 60 features.

3 ECoS for Dynamic Modelling and Classification

Here we use an implementation of the ECoS models called Evolving Classifier Function (ECF) [1]. The ECF algorithm classifies input data into a number of classes and finds their class centres in the n-dimensional input space by "placing" a rule node in the evolving layer. Each rule node is associated with a class and an influence (receptive) field representing a part of the n-dimensional space around the rule node. Generally such an influence field in the n-dimensional space is a hyper-sphere. Essentially each client is modelled by a number of rule nodes that represent that client.

There are two distinct modes of ECF operation, learning and recognition. The details of the original algorithm of these two operation modes were introduced in [1]. In this paper, the recognition algorithm of ECF was modified for the task of person verification. Accordingly we call it verification algorithm. The verification algorithm consists of the following steps:

- With the trained ECF module, when a new test sample I is presented, first it is checked whether it falls within the influence field of the rule nodes representing the claimed identity of the sample I . This is achieved by calculating the Euclidean distance between this sample and appropriate rule nodes, then comparing this distance D_i with the corresponding influence field Inf_i. The sample I is verified as person i if the relation (1) is satisfied.

$$D_i <= \mathrm{Inf}_i \tag{1}$$

- If the sample I doesn't fall in the influence field of any existing rule node,

- Find the rule node which has the shortest distance to this sample, note this distance as D_{min}.
- If this distance D_{min} is less than a pre-set acceptance threshold θ, the sample I is verified as person i. Otherwise, this sample is rejected by this verification module.

This verification algorithm was applied to the speaker, face image and integrated verification modules. Figure 1 illustrates the overall process of adaptive person verification system.

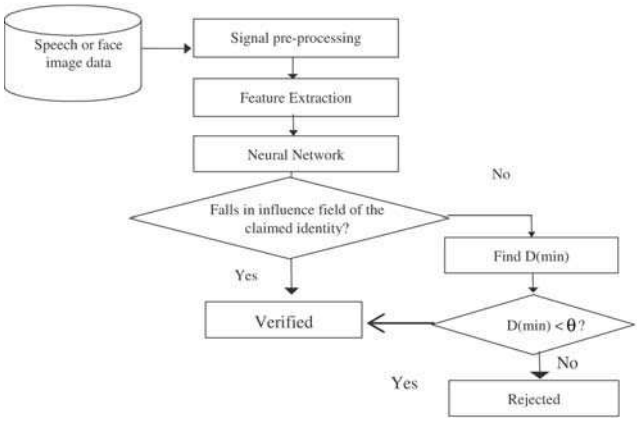

Fig. 1. Overall view of an adaptive connectionist person verification system

4 Integration of Speech and Face Image at the Feature Level

Speech and face image information were used for the person verification task. Individual ECF modules were built for both speech and face image sub-network. In addition, features obtained from speech and face image of clients were merged to form integrated input features. There are various strategies of combining multimodal sources of information. In this approach, speech and face image information were integrated at the feature level. There are 100 input features in a speech sample and 64 input features in a face image sample. These two set of features were concatenated to form the integrated input features.

5 System Implementation and Experimental Results

Person verification system is essentially a two-class decision task where the system can make two types of errors. The first error is a false acceptance, where an impostor is accepted. The second error is false rejection, where a true claimant is rejected. False Acceptance Rate (FAR) and False Rejection Rate (FRR) are calculated according to the following equations:

$$FAR = \frac{I_A}{I_T} \quad (1) \qquad FRR = \frac{C_R}{C_T} \quad (2)$$

where I_A the number of impostors classified as true claimants, I_T is the total number of impostor presented, C_R is the number of true claimant classified as impostors and C_T is the total number of true claimant presented. The trades off between these errors are adjusted using the acceptance threshold θ .

5.1 Data Preparation

In this study, speech data were taken from 8 members of the KEDRI institute [3]. As the speech module is text-dependent, all the speakers were requested to say the word "security" for speech-based speaker verification. Five samples from each speaker were collected to form the training dataset. Another 5 samples from each of these speakers were used to form a testing dataset. In a similar fashion the face images of the same peoples were captured to prepare training and testing datasets. Finally, the input features from speech and face image were integrated according to the method described in section 4.

5.2 Experiments and Results

An Adaptive Speaker Verification Module. An ECF neural network engine was built based on the speech training dataset. Each speaker was modelled by allocating rule nodes during training session. The number of rule nodes assigned for each speaker is determined by the maximum influence field. Figure 2 illustrates the performance of ECF on testing dataset.

As shown in Figure 2, the smaller the Maximum Inference Field, the more rule nodes are allocated for each client. This leads to a high correct acceptance rate of 92% and small FRR and FAR errors of 1%.

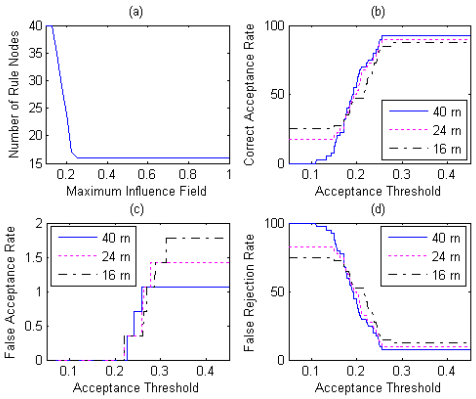

Fig. 2. ECF performance on speaker verification task. (a) Number of rule nodes created vs various influence field values. (b) Correct acceptance rate vs acceptance thresholds. (c) FAR vs acceptance thresholds. (d) FRR vs acceptance thresholds

An Adaptive Face Image Verification Module. Image verification system was built and validated. In a similar fashion to speaker verification model, each person was

modelled by allocating rule nodes during training session. The number of rule nodes assigned for each client is determined by the maximum influence field. Figure 3 illustrates the performance of ECF on testing dataset.

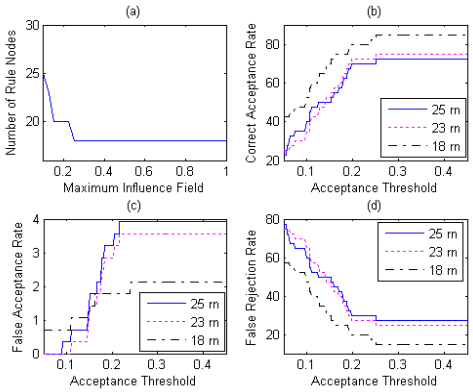

Fig. 3. ECF performance on face image verification task. (a) Number of rule nodes created vs various influence field values. (b) Correct acceptance rate vs acceptance thresholds. (c) FAR vs acceptance thresholds. (d) FRR vs acceptance thresholds

As illustrated in Figure 3, the smaller the Maximum Inference Field, the more rule nodes are allocated for each person. The best ECF performance was achieved with 18 rule nodes with the correct acceptance rate of 85% and FAR error of just over 2%.

A Person Verification Module Based on Integrated Voice and Face Features. The training dataset for this experiments obtained by concatenating the speech training dataset and face image training dataset as described in Section 4. Each integrated sample has 164 input features. An ECF model was built using the integrated training dataset and test on the integrated testing dataset A and B. The test results are shown in Figure 4.

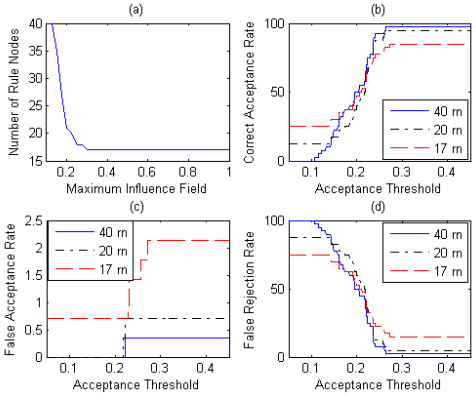

Fig. 4. ECF performance on integrated features. (a) Number of rule nodes created vs various influence field values. (b) Correct acceptance rate vs acceptance thresholds. (c) FAR vs acceptance thresholds. (d) FRR vs acceptance thresholds

The results in figure 4 show that the smaller the Maximum Inference Field, the more rule nodes are allocated for each person. The best ECF performance was achieved with correct acceptance rate of 97% and FAR error of just less than 0.5%.

6 Conclusions and Future Research

This paper presented a method based on evolving connectionist system ECF for person verification tasks. The performance of ECF of individual and integrated modules shows the ECF capability in modelling each person by placing rule nodes to create the person's verification model. The verification module based on integrated voice and face features outperformed both single modules showing improvement in correct acceptance rate and lower FAR and FRR errors. The evolving property of ECF [1] allows for new persons to be added or removed from the system. Further experiments and analysis need to be done to evaluate the performance of this methodology on persons who are not participated in training.

References

1. Kasabov N.: Evolving connectionist systems: Methods and applications in bioinformatics, brain study and intelligent machines, Springer Verlag, 2002
2. Ghobakhlou A., Watts M. and Kasabov N.: Adaptive speech recognition with evolving connectionist systems, Information Sciences 156(2003), 71-83
3. Knowledge Engineering & Discovery Research Institute, Auckland University of Technology, New Zealand, http://www.kedri.info
4. Kasabov N., Postma E.,Herik J. V. D.: AVIS: a connectionist-based framework for integrated auditory and visual information processing, Information Sciences 123 (2000), 127-148
5. Brunelli R., Falavigna D.: Person identification using multiple cues, IEEE Transactions on Pattern Analysis and Machine Intelligence 17(1995), 955-966
6. Luettin J., Thacker N. A., Beet S.W.: Active shape models for visual speech feature extraction, in: D.G. Storck, M.E.Heeneeke(Eds.), Speechreading by Humans and Machines, Springer, Berlin, 1996, 383-390