2004 8th International Conference on
Control, Automation, Robotics and Vision
Kunming, China, 6-9th December 2004

An Adaptive Model of Person Identification Combining Speech and Image Information

David Zhang[1], Akbar Ghobakhlou[1] and Nikola Kasabov[1]

1. Knowledge Engineering and Discovery Research Institute - KEDRI, Auckland University of Technology, Auckland, New Zealand, dzhang@aut.ac.nz , aghobakhlou@aut.ac.nz , nbasabov@aut.ac.nz

Abstract

The paper introduces a combination of adaptive neural network systems and statistical method for integrating speech and face image information for person identification. The method allows for the development of models of persons and their on-going adjustment based on new speech and face images. The method is illustrated with a modeling and classification of different persons, when speech and face images are presented in an incremental way. In this model, there are two sub–networks, one for face image and one for speaker recognition. A higher-level layer is applied to make a final decision. In the speaker recognition sub-network, a text-dependant model is built using Evolving Connectionist Systems (ECOS) [1]. In the face image recognition sub-network, composite profile technique is applied for face image feature extraction and Zero Instruction Set Computing (ZISC) [2] technology is used to build the neural network. In the higher-level conceptual subsystem, final recognition decision is made using statistical method. The experiments show that ECOS and ZISC are appropriate techniques for the creation of evolving models for the task of speaker and face recognition individually. It is also shown that the integration of the speech and image information using statistical method improves the person identification rate.

1. Introduction

Automatic detection of person identity based on biometrics is a commercially very important problem. It arises in security and surveillance applications where access to services, buildings or files should be restricted to authorised individuals. Many low risk applications of the technology also exist, such as the retrieval of faces from video and image databases, video annotation, computer logging, mobile phone security and countless others.

There are many biometric features that distinguish individuals from each other, thus many different sensing modalities have been developed [6]. The most widely used feature is the fingerprint, as there are very cheap sensors that can acquire finger print signatures. Other popular modalities include face and voice as a normal means of interaction between human and machine. More sophisticated sensing techniques exploit the unique pattern of the iris or the thermal signature of the human face acquired by infrared camera. These can be used successfully individually, as exemplified by the iris scan system deployed in the banking sector and currently being tested for airport security [7].

Over the last 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. This is particularly advantageous when the system combines relatively weak or fragile modalities such as voice and face images. Although speaker identification using clear speech is very effective and reliable, it degrades rapidly in noisy environments. Similarly, face recognition and identification is seriously affected by lighting conditions and by variations in the subject's pose in front of the camera. The advantage of multimodal approaches is that the resulting systems are likely to be more robust for environmental conditions. Moreover, in good conditions their joint use should lead to significantly better recognition and identification performance than can be achieved with single modality systems.

In conventional neural network, any re-training process will modify the connection weights of a static structure of the neural network. This often leads to the problem of forgetting the previous knowledge. From previous work [3], it was shown that an evolving connectionist system (ECOS) can be used to create model for adaptive speech recognition system. Adaptation is a process of accommodating new instances of data (in this paper, image and speech) that were mis-recognized by the system. Both ECOS and Zero Instruction Set Computing (ZISC) neural network use local learning, each neuron in the evolving layer of the neuron network represents a small region (area) in the problem domain. Both ECOS and ZISC are adaptive connectionists that allow a structural modification.

0-7803-8653-1/04/$20.00 © 2004 IEEE

In this paper, two sources of information, speech and face image were chosen for the task of person identification.

Research questions that are attempted in the paper are:

- Can an ECOS be used to create a model for speaker recognition? How is the recognition rate? Is the model adaptive?
- Can a Zero Instruction Set Computing (ZISC) be used to create a model for person identification using face image? How is the performance? Is the model adaptive?
- Does integrating the information from speaker recognition and face image recognition engine improve the person identification rate?

2. Speech and face image signal processing

2.1 Speech signal sampling and processing

In the speaker recognition system, 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. Spectral analysis of the speech signal was performed over 20ms with Hamming window and 50% overlap, in order to extract mel frequency cepstrum coefficients (MFCC) as acoustic features. Discrete cosine transformation (DCT) was applied on the MFCC of the whole word in the following manner.

For an m frame segment, DCT transformation will result in a set of m DCT coefficients. This sequence is truncated to achieve a fixed-size input vector consisting of 20×d, where d is the dimensionality of the feature space [3]. Figure 1 illustrates the feature extraction procedure used to obtain input feature vectors.

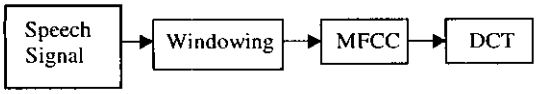

Fig 1: Feature extraction procedure

2.2 Face image processing

In the face recognition sub-system, 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.

As the length of the features is 560 (320+240), this exceeds the maximum vector length of 64 bytes supported by the ZISC neural network (extended from the ZISC36 chip), the features were mapped to 64 bytes by interpolation.

3. ECOS for dynamic modeling and classification in speech sub-network

Here we use an implementation of the ECOS models called Evolving Classifier Function (ECF) [1]. The ECF algorithm outlined below, classifies a data set into a number of classes and finds their class centres in the n-dimensional input space by "placing" a rule node. 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.

There are two distinct modes of ECF operation, learning and recognition. During the learning mode, data vectors are fed into the system one by one with their known classes. The learning sequence of each iteration is described as the following steps:

1) If all vectors have been inputted, finish the current iteration; otherwise, input a vector from the data set and calculate the distances between the vector and all rule nodes already created;
2) If all distances are greater than a max-radius parameter, a new rule node is created. The position of the new rule node is the same as the current vector in the input data space. Its radius is set to the min-radius parameter, go to step 1; otherwise:
3) If there is a rule node with a distance to the current input vector less then or equal to its radius and its class is the same as the class of the new vector, nothing will be changed and go to step 1; otherwise:
4) If there is a rule node with a distance to the input vector less then or equal to its radius and its class is different from those of the input vector, its influence field should be reduced. The radius of the new field is set to the larger value from the distance minus the min-radius, and the min-radius.
5) If there is a rule node with a distance to the input vector less then or equal to the max-radius, and its class is the same to the vector's, enlarge the influence field by taking the distance as the new radius if only such enlarged field does not cover any other rule node which has the different class; otherwise, create a new rule node the same way as in step 2, and go to step 1.

The recognition is performed in the following way:
1) If the new input vector lies within the field of one or more rule nodes associated with one class, the vector belongs to this class;
2) If the input vector lies within the fields of two or more rule nodes associated with different classes, the vector will belong to the class corresponding to the closest rule node.
3) If the input vector does not lie within any field, then there are two cases: (1) one-of-n mode: the vector will belong to the class corresponding to the closest rule node; (2) m-of-n mode: select m rule nodes that generate the highest activation for the new vector. Find

the distances between this vector and the selected nodes. Then calculate the average distance according to each class. The vector belongs to the class corresponding the smallest average distance.

In previous work [3], it is demonstrated that ECOS is an efficient tool for building adaptive speech recognition systems. Can it also be used to build adaptive speaker recognition systems? Figure 2 illustrates the overall view of an adaptive speaker recognition system.

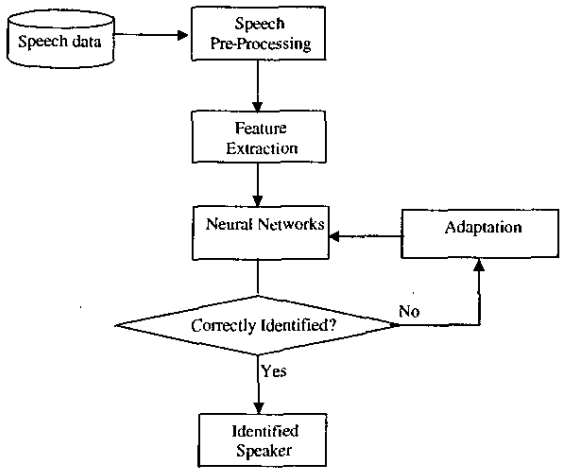

Fig 2: Flowchart of adaptive speaker recognition system

4. Modified ZISC for dynamic modeling and classification in image sub-network

Zero Instruction Set Computing (ZISC) can be considered as an expert system that can recognize and classify objects or situations and take instantaneous decisions based upon accumulated knowledge. A ZISC engine is built on a model that consists of mapping an N-dimensional space by rule nodes. Each rule node is associated with category and an influence field representing a space around the rule node, where generalization is possible. The classification of a new vector requires the calculation of the distances between the input vector and the rule nodes stored in the knowledge base. The engine sorts the distances using K-Nearest Neighbors (KNN) mode.

Again, there are two modes in a ZISC engine, learning and recognition mode.

The learning process requires feeding to the network a set of vectors with their known category. It can result in the following actions:

- If the vector does not fall in any of the influence fields of the rule nodes already stored in the network, a new neuron is committed. Its influence field is set to the minimum value between the Maximum Influence Field (MaxIF) and the distance to the closest rule node.
- If the vector falls in the influence field of a rule node already stored in the network and their category matches, no change to the network.
- If the vector falls in the influence field of a rule node already stored in the network but their category does not match, the action will be:
  → One or more influence fields are reduced so the adjacent neurons with different categories become tangent. The reduction of the influence field however cannot go beyond a minimum defined by the Minimum Influence Field (MinIF, another global parameter).
  → If the reduction of the influence field is decreased to the MinIF, the neuron is labeled as "degenerated".

In the recognition mode, the network decision is taken upon the result of the following comparisons:

- If the input vector does not lie within any influence fields, it is not recognized.
- If the input vector lies within the influence field of one or more rule nodes associated with one category, it is recognized and declared as belonging to this category.
- If the input vector lies within the influence field of two or more rule nodes associated with different categories, it is declared as unidentified, that is recognized but not formally identified.

A modification was made to the ZISC engine recognition mode to calculate activation for each rule node. Let's assume a network with N rule nodes R_i (i=1,2,...,N) and M different categories C_i (i=1,2,...,M) (one for each person). The normalized Hamming distance for a new vector V is calculated according to Equation 1 and N distances D_i, (i=1,2,...,N) are calculated.

$$D_i = \frac{\sum_k^L |V_k - R_{ik}|}{\sum_k^L |V_k + R_{ik}|} \qquad i=1,2,....,N \qquad (1)$$

Where, L is the number of input features, V_k is the k-th feature of vector V and R_{ik} is the k-th feature of rule node R_i.

Then, activation of each rule node is calculated according to Equation 2 using a linear activation function. Other activation functions, such as a radial basis function could be used.

$$A_i = 1 - D_i \qquad i=1,2,....,N \qquad (2)$$

As each rule node relates to one category, the N activations can be divided into M sets, each set relates to only one category. Then, choose the maximum activation of each set (accordingly, each category) and create a vector by putting these activations together.

Figure 3 shows the flowchart for the image recognition module.

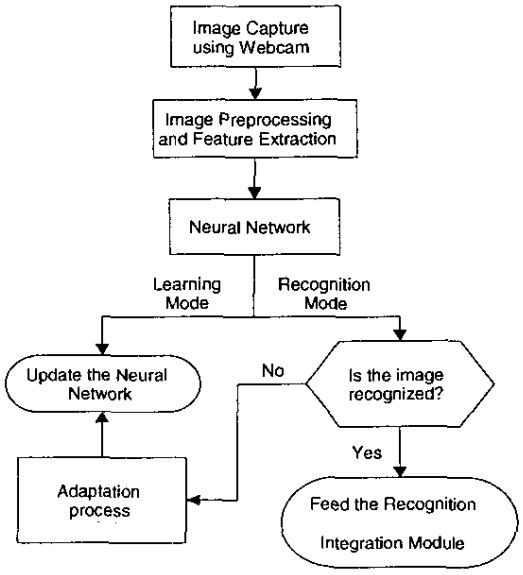

Fig 3: The image recognition module

5. The higher-level conceptual subsystem using statistical method

The higher-level conceptual layer takes its inputs (the activations) from both speech and face recognition sub-network and makes a decision on the identity of the person observed. There are various strategies of combining multimodal sources of information. In this paper, we use the principle of statistically based specialization for taking decisions based on different sources of information [5].

In general, the speech and face recognition subsystems deal with different parts of task. For instance, the speech subsystem is responsible for recognizing a person's voice and the image subsystem for recognizing a person's face. Each of the subsystems makes its own contribution to the overall task. The conceptual subsystem weights the contributions of the two subsystems according to their individual recognition rates.

The method of assigning weights for the contribution from image and speaker recognition modules is computed in the following manner. If the recognition probability of the image subsystem and speech subsystem for the output category j(person j)is $P_{image,j}$ and $P_{speech,j}$, then the weights of the two inputs to the conceptual subsystem $W_{image,j}$ and $W_{speejh,j}$ can be calculated using Equation 3.

$$W_{image,j} = \frac{P_{image,j}}{P_{image,j} + P_{speech,j}}$$

$$W_{speech,j} = \frac{P_{speech,j}}{P_{image,j} + P_{speech,j}} \qquad (3)$$

Where $P_{image,j}$ is the recognition rate for the j-th person of the face recognition neural network and $P_{speech,j}$ is the recognition rate for the j-th person of the speaker recognition neural network.

The higher-level conceptual layer takes the activation output vectors from both image and speech sub-network (noted as $A_{image,j}$ and $A_{speech,j}$.) as inputs. The final decision of person identity is made according to the following steps:

a. Calculate the overall activation set $A_{final,j}$ according to Equation 4

$$A_{final,j} = W_{image,j} \times A_{image,j} + W_{speech,j} \times A_{speech,j} \quad (4)$$

b. Find the element with maximum activation in the overall activation set A_{final}.

c. Compare this activation value with a pre-set threshold θ

- If it is larger than or equal to θ, the category (person) related with this activation is announced to be the recognition result.
- Otherwise, the conceptual sub-network declares the current test sample to be "Unknown Person", none of the person known by the neural networks is recognized.

Figure 4 shows the flowchart for the integration process.

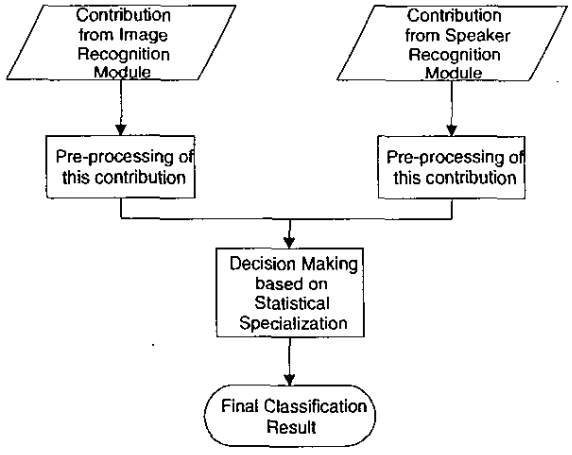

Fig 4: The flowchart for the integration process

6. System Implementation and Experimental results

6.1. Adaptive speaker recognition module and its validation

An adaptive experimental speaker identification system based on speech input was built. Speech data were taken from 7 members of the KEDRI institute [4]. As the speech module is text-dependent, all the speakers

were requested to say the word "security" for speech-based speaker identification.

The experiments were designed in two distinct phases. In the first phase, a neural network engine was built to recognize 5 speakers (person A to E). Training dataset contains 50 samples (10 samples from each speaker) and testing dataset contains 100 samples (20 samples from each speaker).

The second phase of this experiment was designed to evaluate the adaptation ability of this recognition model by enrolling new speakers (person F and G). Two new speakers were added incrementally. 10 samples for each person were used for furthering training and another 20 samples for each person were used as testing dataset.

The recognition accuracy of each person before and after adaptation is shown in Table 1.

Table 1: Speaker recognition accuracy before and after adaptation

Person	Recognition Accuracy before adaptation (%)	Recognition Accuracy after the adaptation on person F (%)	Recognition Accuracy after the adaptation on person G (%)
A	90	90	90
B	80	80	80
C	70	70	70
D	85	85	85
E	90	90	90
F	N/A	85	85
G	N/A	N/A	95
Average	83	83.33	85

An average recognition accuracy of 83% was obtained for the first 5 speakers (A to E). Table 1 also shows the performance of ECF after adding person F and G. As illustrated, while the engine was expanded to be able to recognize 2 additional speakers, it maintains the performance on the pervious speakers. It demonstrates that ECF is an efficient neural network for building a model for speaker recognition, and more important, it is adaptive.

6.2 A face recognition module using ZISC neural network and its validation

An adaptive face image recognition system was built and validated. Face image from the same 7 members of KEDRI group were involved.

Similar to the process of creating a speaker recognition model, the experiments were designed in two distinct phases. In the first phase, a neural network engine was built to recognize 5 persons (A to E). Training dataset contains 50 samples (10 samples from each person) and testing dataset contains 100 samples (20 samples from each person).

The second phase of this experiment was designed to evaluate the adaptation ability of this recognition model by enrolling new persons (person F and G). Two new persons were added incrementally. 10 samples for each person were used for furthering training and another 20 samples for each person were used as testing dataset.

The recognition accuracy of each person before and after adaptation is shown in Table 2.

Table 2: Face recognition accuracy before and after adaptation

Person	Recognition Accuracy before adaptation (%)	Recognition Accuracy after the adaptation on person F (%)	Recognition Accuracy after the adaptation on person G (%)
A	75	75	75
B	80	80	80
C	65	65	65
D	90	90	90
E	80	80	80
F	N/A	85	85
G	N/A	N/A	80
Average	78	79.17	79.29

An average recognition accuracy of 78% was obtained for the first 5 persons (A to E). Table 2 also shows the performance of ZISC after adding person F and G. As illustrated, while the engine was expanded to be able to recognize 2 additional persons, it maintains the performance on the pervious persons. The results demonstrate that ZISC is an efficient neural network for building an adaptive model for face recognition.

6.3 Person identification by integrating the outputs from the two modules built above

The recognition rates of each of the speech and face modules for each person (Tables 1 and 2) were used in Equation 3 to calculate the integration weights W_{speech} and W_{image}. When a pair of speech and image sample data of a person were entered in the system (face image data to image sub-network, and speech data to speech sub-network), the activation outputs A_{image} and A_{speech} were calculated and then fed into the higher-level conceptual subsystem as inputs. The final recognition result can be calculated using Equation 4 combining with the threshold value θ (within this experiment, θ was set to 0.87).

The performance of speaker recognition model, face recognition model and the integration model are shown in Table 3.

Table 3: Performance of the speaker recognition model, face recognition engine and integration model

Person	Speaker recognition model (%)	Face recognition model (%)	Integration Model (%)
A	90	75	90
B	80	80	90
C	70	65	80
D	85	90	95
E	90	80	95
F	85	85	90
G	95	80	100
Average	85	79.29	91.43

The average recognition accuracy is 91.43%. The result shows that by integrating the speaker and face recognition engine as proposed in [5], the average recognition accuracy is enhanced, higher than the speaker recognition accuracy (85%) and face image recognition accuracy (79.29%). Figure 5 below shows the false acceptance rate (FAR) and falsely rejection rate (FRR) of the person identification system. The horizontal axis is the threshold varying from 0.6 to 0.95, with a step 0.001. The dotted curve represents FAR, solid curve for FRR. The point where these two curves cross each other is the equal error rate (EER).

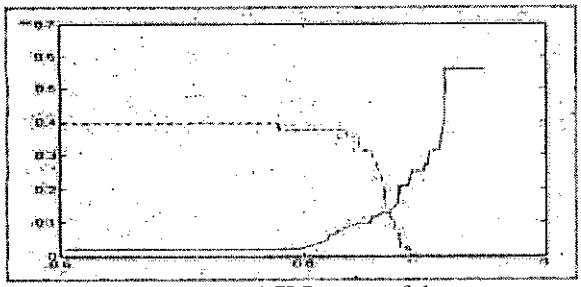

Fig 5: the FAR and FRR curve of the person identification system according to different threshold

7. Summary and Conclusions

This research and experiments show that ECOS [1] and ZISC [2] are appropriate techniques for the creation of models for the task of speaker and face recognition individually. The models are adaptive, which means the recognition accuracy of these models for the existing persons can be improved through efficient adaptation. The models can be expanded to accommodate new classes (persons) without degrading its performance over the existing classes. The experiments also show that by integrating the speech and image information using the statistical method from [5], the person identification rate was improved.

In this paper, statistical method is applied for integrating the speech and face image information. Future work can be extended in the direction of finding a novel connectionist-based method for this integration process.

Acknowledgements

The research is funded by the FRST/NERF grant NERF/AUTX02-001.The speech and face image data were collected in the Knowledge Engineering & Discovery Research Institute KEDRI [4].

References

[1]. N. Kasabov, Evolving connectionist systems: Methods and applications in bioinformatics, brain study and intelligent machines, Springer Verlag, 2002

[2]. ZISC Manual, Zero Instruction Set Computing (ZISC), 2000. Available from http://www.silirec.com

[3]. A. Ghobakhlou, M. Watts and N. Kasabov, "Adaptive speech recognition with evolving connectionist systems", Information Sciences 156(2003), pp 71-83

[4.] Knowledge Engineering & Discovery Research Institute, Auckland University of Technology, New Zealand, http://www.kedri.info

[5]. N. Kasabov, E. Postma, J. V. D. Herik, "AVIS: a connectionist-based framework for integrated auditory and visual information processing", Information Sciences 123 (2000), 127-148

[6]. R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Transactions on Pattern Analysis and Machine Intelligence 17(1995), 955-966

[7]. J.Luettin, N.A.Thacker, S.W.Beet, Active shape models for visual speech feature extraction, in: D.G. Storck, M.E.Heeneeke (Eds.), Speechreading by Humans and Machines, Springer, Berlin, 1996, pp. 383-390

Appendix

On the basis of the integration method in this paper, a software environment for person identification has been developed. Below is a snapshot of the interface of this software environment. For more information, please visit our website at http://www.kedri.info, or contact the authors: dzhang/nkasabov/akbar@aut.ac.nz