

Chapter 13

Evolving Systems for Integrated Multimodal Information Processing

Nik Kasabov

nkasabov@aut.ac.nz

www.kedri.info

Overview

- Multi-modal information processing
- AVIS: Evolving Connectionist Framework for Integrated Auditory and Visual Information Processing Systems
- PIAVI - An Experimental Evolving System for Person Identification based on Integrated Auditory and Visual Information Processing

Multimodal Information Processing

- Many processes of perception and cognition are multimodal, involving auditory-, visual-, tactile-, and other type of information processing.
- Processes are extremely difficult to model without having a flexible, multi-modular evolving system in place
- Information from different modalities can support the performance of a computer system originally designed for a task with a unimodal nature
- Image information, auditory information, and textual input can be used to enhance the recognition of objects (e.g. the identification of moving objects based on information coming from their blurred images and their sounds)

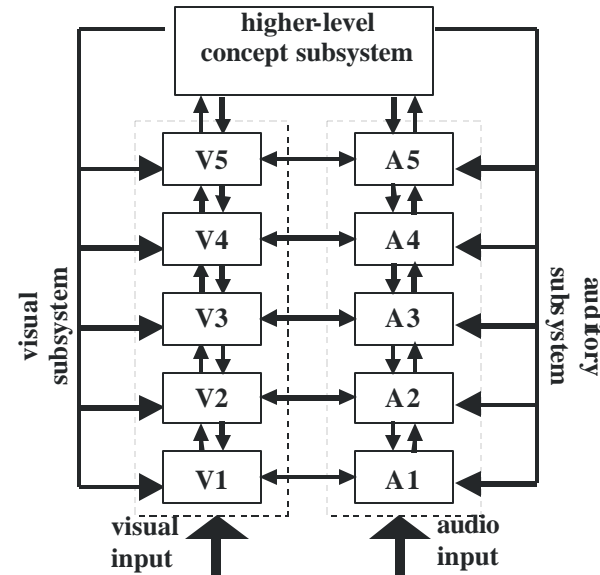
Multimodal Information Processing

- Integrating auditory and visual information in one system requires the following four questions to be addressed.
 - » Auditory and visual information processing are both multilevel and hierarchical (ranging from an elementary feature extraction level up to a conceptual level). So, at which level and to what degree should the two information processes be integrated?
 - » How should time be represented in an integrated audio-visual information processing system? This is a problem related to the synchronisation of two flows of information. There could be different scales of integration, e.g., milliseconds, seconds, minutes, hours.
 - » How should adaptive learning be realised in an integrated audiovisual information-processing system? Should the system adapt each of its modules dependent on the information processing in the other modalities?
 - » How should new knowledge (e.g., new rules) be acquired about the auditory and the visual inputs of the real world?

AVIS

- Three subsystems:
 - » Auditory
 - » Visual
 - » Higher-level conceptual
- AVIS allows an auditory subsystem, as well as a visual subsystem, to operate either as separate subsystems, or together.
- AVIS can operate in six main modes:
 - » Unimodal auditory
 - » Cross-modal auditory
 - » Bimodal auditory
 - » Unimodal visual
 - » Cross-modal visual
 - » Bimodal visual

AVIS



A block diagram of a framework for Auditory and Visual Information Processing Systems (fig. 13.1)

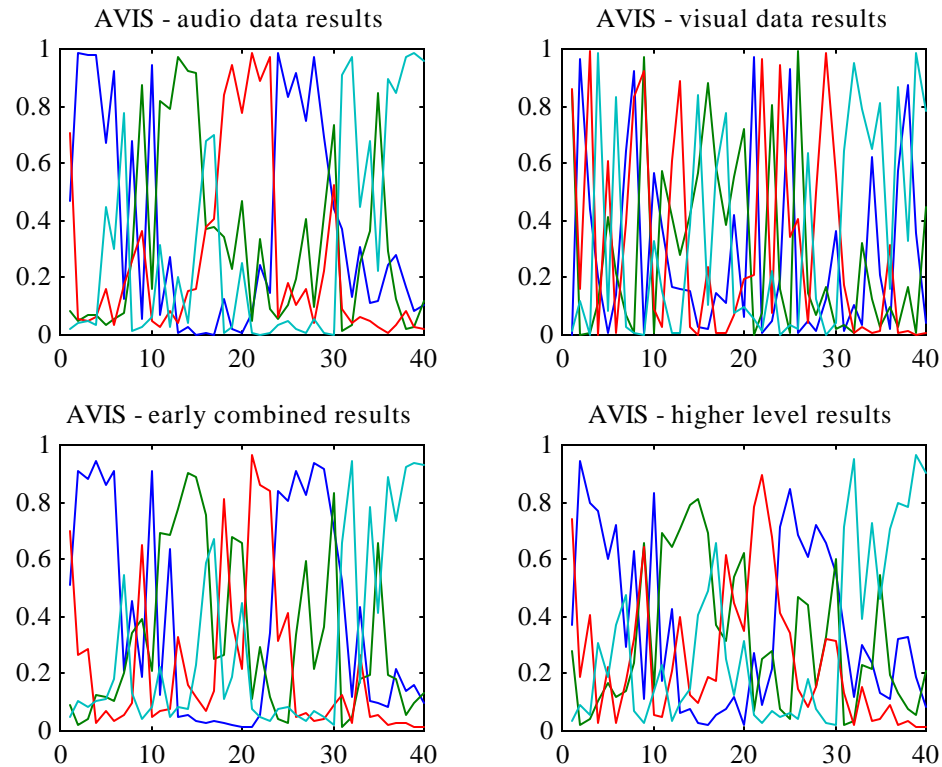
PIAVI – Experimental Evolving System

- Modelled on the AVIS structure, although the auditory and visual subsystems each consist of single modules
- Each of the subsystems is responsible for a modality-specific subtask of the person-identification task
- PIAVI has four modes of operation
 - » Unimodal visual mode
 - » Unimodal auditory mode
 - » Bimodal mode
 - » Combined mode

PIAVI – Case Study 1

- Aims at evaluating the added value of combining auditory and visual signals in a person-identification task.
- Our emphasis on dynamical aspects implies that the integration of auditory and visual information requires an extended period of time
- The images (i.e., frames of video data) contained in each segment need to be transformed into a representation of the spatio-temporal dynamics of a person's head
- The generalisation performance is defined as the classification performance on the test set.
- The overall recognition rate achieved in the combined mode is
 - » 22% higher than the recognition rate in the unimodal visual mode,
 - » 17% higher than the recognition rate in the unimodal auditory mode, and
 - » 4% higher than the recognition rate in the early-integration mode of operation.

PIAVI – Case Study 1

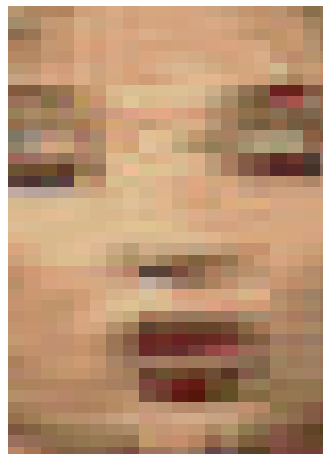


The results obtained. The test frames are shown on the x-axis, the output activation values are shown on the y-axis. (Fig. 13.2)

PIAVI – Case Study 2

- Attempts to improve and extend the results obtained in the first case study by employing a larger data set and by defining aggregate features representing longer temporal intervals
- Used CNN broadcasts of eight fully visible and audibly-speaking presenters of sport and news programs
- The unimodal visual mode of operation is modelled in an EFuNN with 60 (N+M) input nodes, 180 input membership functions and 16 output membership functions.
- The bimodal mode of operation is modelled using an EFuNN with 86 input nodes and 258 membership functions
- Combination of the visual and auditory data using bimodal processing, yielded a major improvement in generalisation performance

PIAVI – Case Study 2



Person	generalisation performance (%)	
	unimodal	bimodal
1	100	100
2	28	88
3	92	96
4	72	80
5	60	64
6	48	96
7	100	100
8	96	100

Left – An example of a frame used in the data set (Fig. 13.3)

Middle – The face template used for video frames applied to the speaker video data from the above figure (Fig. 13.4)

Right – Generalisation performances for unimodal (visual data) and bimodal (visual and auditory data) experiments (Table 13.1)

Discussion

- Results of case studies prove added value of integrating auditory and visual information for person identification.
- In Case Study 1 combining the auditory and visual information enhanced the generalisation performance.
- Case Study 2 showed that, with a large number of training examples, unimodal processing on the basis of dynamic visual features leads to a perfect performance on a large data set.
- From a practical viewpoint, the use of smaller training sets, facilitates the speed at which the PIAVI system in its bimodal mode of operation learns to classify persons from video data.

Summary

- Introduced AVIS framework, which facilitates the study of:
 - » different types of interaction between modules from hierarchically-organised subsystems for auditory and visual information processing;
 - » early and late integration of the auditory and the visual information flows, dynamic auditory and visual features;
 - » pure connectionist implementations at different levels of information processing; and
 - » evolving fuzzy neural networks that allow for learning, adaptation, and rule extraction.
- The integrated processing of auditory and visual information may yield:
 - » an improved performance on classification tasks involving information from both modalities
 - » reduced recognition latencies on these tasks
- The AVIS framework has a potential for many applications for solving difficult AI problems.

Further Readings

- Sources of neural structure in speech and language processing (Stork, 1991).
- Integrating audio and video information at different levels (Waibel, et al 1997).
- Using lip movement for speech recognition (Stork and Hennecke, 1996; Massaro and Cohen, 1993; Gray et al, 1997; Luettin, J., 1996).