Audio-Visual Localisation and Tracking Using Neurobiologically-Inspired Techniques*

Stuart N. Wrigley and Guy J. Brown

Speech and Hearing Research Group, Department of Computer Science,
University of Sheffield, UK
s.wrigley@dcs.shef.ac.uk, g.brown@dcs.shef.ac.uk

1 Extended Abstract

Perception is a process which creates a mental representation of the world from the multitude of information gathered by the senses. In order to produce a representation of an object, the brain must provide a solution to the *binding problem*: how does the brain, confronted with many features, encoded in many different regions, draw them all together to form a perceptual whole? This problem arises in regard to feature combination within a single modality (e. g. the binding of edges, textures and colours to form a visual image). However, the binding problem also concerns the broader issue of how to link features in *different* modalities, such as the association of a sound with a visual object and possibly even a smell.

A solution to the binding problem employs a large number of spatially distributed neurons in which an individual neuron can be a member of several assemblies (each representing a different perceptual object). von der Malsburg [1] suggests that different assemblies could be distinguished by temporal synchronisation of the responses of their constituent neurons — the *temporal correlation* theory of binding. A successful implementation of the temporal correlation theory uses oscillators to represent the mean discharge response of a pool of neurons. Groups of features form wholes if their associated oscillators are synchronised and the oscillations of unrelated wholes are desynchronised.

Although a large number of computational studies have demonstrated the use of neural oscillators for segregation and grouping of objects within a single modality (for a review see [2]), few have examined their utility in computational models of across-modality binding. We investigate the use of neural oscillators for audio-visual grouping using a localisation and tracking problem. The goal of the system is to determine the spatial location of an individual participant and track that participant through time.

Audio information is acquired from a KEMAR binaural manikin and visual cues from a single camera, placed directly above the manikin. An estimate of the sound source's azimuth is calculated by cross-correlating the output of an auditory peripheral hearing model. The visual features consist of binary masks representing the locations within a frame of faces; objects which are not usually present in the room (these tend to be participants) and motion.

The audio visual grouping mechanism operates on the basis of common spatial location. The video and audio features are represented on two locally excitatory globally inhibitory networks of relaxation oscillators (LEGION, [3]).

* Presented at 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Edinburgh, UK, 11-13 July 2005

Cross-modal segmentation occurs due to the placement of excitatory connections between the two networks. These connections are placed between azimuth nodes and video columns and the strengths are established via an initial training phase for the meeting room in which activity in a particular region of the video frame is mapped to audio activity at a particular azimuth. The nature of the networks allows multiple participants to be represented simultaneously.

Following the integration of audio and video features, an individual A-V object is selected and tracked using a model based upon the neurobiological and psychophysical behaviour of the human oculomotor system. Specifically, smooth pursuit and saccadic eye movements are used to track an object based on its velocity.

The oculomotor model is evaluated on a number of trials based upon the Rashbass paradigm in which a target step in one direction is immediately followed by constant velocity in the other [4]. In addition, a second step is included followed by a constant velocity in the same direction. The response of the model is compared against published human behavioural data for the same trials [5] and displays a high qualitative match.

The overall A-V tracking system is evaluated using a recording of a single participant who moved around the meeting room uttering a short phrase at 10 degree intervals. The system tracks the participant with high accuracy; the mean absolute error per frame across the entire sequence is only 13.1 pixels — much less than the width of a face (26 to 46 pixels depending on the distance from the camera).

2 Acknowledgment

This work was conducted as part of the MultiModal Meeting Manager (M4) and Augmented Multi-party Interaction (AMI) projects, funded by the EU IST Programme. The authors thank Darren Moore, IDIAP, for his assistance during data collection.

References

1. von der Malsburg, C.: The correlation theory of brain function. Technical Report 81-2, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany (1981)
2. Wang, D.L.: The time dimension for scene analysis. IEEE Trans. Neural Networks (in press)
3. Terman, D., Wang, D.L.: Global competition and local cooperation in a network of neural oscillators. Physica D **81** (1995) 148–176
4. Rashbass, C.: The relationship between saccadic and smooth tracking eye movements. Journal of Physiology (London) **159** (1961) 326–338
5. de Brouwer, S., Yuksel, D., Blohm, G., Missal, M., Lefèvre, P.: What triggers catch-up saccades during visual tracking. J. Neurophysiol. **87** (2002) 1646–1650