Synergistic Acoustic, Visual, and Motor Interactions

Ed Gamble and David Rainton
ATR Human Information Processing Research Laboratory

2-2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto-fu 619-02 Japan
ebg@hip.atr.co.jp

We describe our ongoing research program to exploit interactions between various sensory inputs, such as visual, speech and motor inputs, to overcome limitations in each of the sensory inputs and to facilitate human-like interaction. The noise processes in each of these inputs are largely orthogonal to each other and therefore significant benefits arise from *inter-sensory* integration of information. We apply these interactions towards our goal of building a system capable of real-time spoken and visual interactions with humans in the limited, yet important, domain of multimedia teleconferencing.

聴覚・視覚・運動の共働的相互作用

Ed Gamble　David Rainton
㈱エイ・ティ・アール　人間情報通信研究所

現在、視覚、聴覚（音声）、運動といった様々な感覚の相互作用の利用に関する研究を進めている。その目的は、それぞれの知覚のもつ限界を克服し、人間の知覚に似た感覚間の相互作用を促進することにある。例えば、こちらの感覚における雑音処理は、互いに独立と考えられ、感覚相互の情報の統合は有益な効果を生み出すものと思われる。リアルタイムの視聴覚情報処理による人間との相互作用を可能とするシステム構築を最終目標として、限られた条件ではあるが重要であるマルチメディア・テレビ会議への感覚相互作用の応用を行う。

1 Introduction

We describe a state-of-the-art computational system which addresses some of the most important issues in human-machine interaction and which, through the interaction of multiple perceptual inputs, offsets fundamental errors in any one of the perceptual modules. Our design benefits from a symbiotic combination of auditory and visual information. The benefits of combining information within a perceptual module (that is *intra-perceptual*) is well know within the computational vision community. Our proposal, however, stands to benefit from *inter-perceptual* information combination.

2 The Human-Machine Interaction

We envisage the human-machine interaction as follows. Consider a multimedia, audio-visual teleconferencing application. At each end of the audio-visual link sits an additional interactive conference "participant" comprising of a robotic system with microphone(s), speech synthesizer, motion-controlled camera(s), imaging device, and computational engine. This robotic system will automatically detect who is talking and focus its attention accordingly; this is, lock its audio beam onto a given speaker thereby suppressing background noise and rotate its camera thereby providing a full face image of the speaker. Once concentrating upon a particular speaker, the system will used combined audio/visual tracking techniques to keep hold of a speaker as she moves within the room.

Additional human-machine interaction abilities include the robotic system's abilities in speech recognition and visual face recognition albeit in a limited context. The conference participants can talk to the system just like to any other member of the conference except for the limited vocabulary. For example, the system could be instructed to zoom in or zoom out upon or to focus upon any person in the conference room. Simple commands to start a video or slide presentation, to maintain both an audio and visual record of the conference (robotic-stenographer), and to recall information from databases are all examples of expected functionality.

With low-level speech processing and face recognition abilities, the system can embody the notion of *video meishi*. If at the start of the teleconferencing session, the conference participants are expected to identify themselves so that the system can develop both facial and aural recognition templates, then, as the conference proceeds, the system will continuously identify speakers by voice and by face. So, just as in a Japanese meeting one lays out ones meishi in the order of people sitting in the room, so too can the robotic system produce a set of meishi superimposed upon the video image to indicate the speakers participating in the conference.

Furthermore, based both on auditory and visual processing the system can automatically identify the number of speaking sources in the room. If there is but one speaker, then a live full face image of that person is

transmitted as that person is tracked moving about the room or within her chair. Thus in the case of two people speaking to each other through the audio-visual link at remote conference sites, each sees a full face image of the other. In other words it is just as if they are talking directly across a single conference table.

The robotic system provides a virtual connection between the conference rooms, effectively giving all participants the same view of the current speakers, no matter where they are situated.

3 Interperceptual Information Processing

As these previous examples of the human-machine interaction indicate, an important aspect of the system is the integration of visual and audio information with each being used to improve the other. Imagine the current problem faced when extracting a voice from within a noisy environment. Based on auditory input alone the task might be impossible without accurate estimates of the spectral densities of both the speaker and noise source. If, however, image processing is used to locate a face or, more simply, a moving object, then a multiple microphone system utilizing auditory phase information can hone in upon the visually interesting source. Thus the auditory signal is easily segmented into spatially distance sources.

Consider for a moment the considerable problem of image segmentation. But now imagine a multiple microphone system in addition to a camera. If the microphone system detects a strong auditory signal 'off to the left,' then the camera can rotate and zoom upon that aurally interesting region or object. Admittedly the accuracy of the auditory signal cannot match the segmentation expected from current image segmentation algorithms; however, general segmentation has proved a difficult task and, in our application and possibly others, only the relatively inaccurate segmentation based on the auditory signal may be necessary.

4 Conclusions

As these examples illustrate, combining visual and auditory information greatly simplifies some previously perplexing segmentation and recognition problems in both image and speech processing. Ultimately, this arises because the noise sources in visual and auditory information are fundamentally orthogonal.