Chapter IV
An Efficient Unification-Based Multimodal Language Processor for Multimodal Input Fusion

Fang Chen
National ICT Australia, Australia & University of Sydney, Australia

Yong Sun
National ICT Australia, Australia & University of Sydney, Australia

ABSTRACT

Multimodal user interaction technology aims at building natural and intuitive interfaces allowing a user to interact with computers in a way similar to human-to-human communication, for example, through speech and gestures. As a critical component in a multimodal user interface, multimodal input fusion explores ways to effectively derive the combined semantic interpretation of user inputs through multiple modalities. Based on state-of-the-art review on multimodal input fusion approaches, this chapter presents a novel approach to multimodal input fusion based on speech and gesture; or speech and eye tracking. It can also be applied for other input modalities and extended to more than two modalities. It is the first time that a powerful combinational categorical grammar is adopted in multimodal input fusion. The effectiveness of the approach has been validated through user experiments, which indicated a low polynomial computational complexity while parsing versatile multimodal input patterns. It is very useful for mobile context. Future trends in multimodal input fusion will be discussed at the end of this chapter.

INTRODUCTION TO MULTIMODAL INPUT FUSION

Multimodal interaction systems allow a user to interact with them by using his/her own natural communication modalities such as pen gesture, text and spoken language, graphical marks etc. After speech and pen gesture were used as input modalities in some researches, other modalities, such as hand gesture and eye gaze, have been addressed as well. To understand a user’s meaning conveyed through his/her use of multiple...
modalities, Multimodal Input Fusion (MMIF) integrates users’ input from different modalities and derives a semantic interpretation from them for a system to act upon. It is an area that deals with various input modalities and finds users’ meaning in a particular instance, application or task. With the development of multimodal interface on mobile devices, multimodal input fusion has to deal with computational complexity, besides the existing issues.

The aim of MMIF is to:

a. Identify the combination of input modalities and multimodal input patterns; and
b. Interpret both unimodal input and multimodal inputs; and
c. Cope with modality alternation for one multimodal utterance and support modality combination (complementary) for one utterance.

There is an important distinction between the aims of MMIF:

- Combining complementary inputs from different modalities, for example, a user touching an object to be deleted on a touch screen and saying ‘delete’.
- Improving the robustness of signal recognition, for example audio-visual speech recognition, or combined gaze and speech input (Zhang, Imamiya, Go & Mao 2004).

The pre-request for successful MMIF includes:

- Fine grained time-stamping of user’s multimodal input (beginning and end); and
- Parallel recognizers for a modality. Based on probabilistic methods, the recognition results from parallel recognizers can be used to produce n-best list; and
- Unified representation schema for output of all recognizers; and
- Time sensitive recognition process.

On the other hand, a successful MMIF should be able to

- Handle input signals that may or may not be temporally overlapped; and
- Disambiguate inputs among different modalities; and
- Deal with simultaneous and/or sequential multimodal inputs; and
- Distribute to allow more computational power.

In a mobile device with a multimodal application, speech can be used as the primary modality because traditional user interface peripherals may not be available or appropriate. The most common instance of such interfaces combines a visual modality (e.g. a display, keyboard, and mouse) with a voice modality (speech). Other modalities such as pen-based input may be used, as well. In the more recent mobile devices, modalities can involve global positioning systems (GPS), and mapping technologies, digital imaging, barcode scanning, radio frequency identification (RFID) and other newer, intelligent technologies. In this kind of applications, users have more options to express their meaning, and tend frequently to switch between interaction modalities owing to a number of external factors such as background noise. Because multiple modalities can be used together to express a meaning, multimodal inputs tend to be more concise and meaningful. For example, “from the corner of Edi Ave. and George Street” can be concisely expressed as “from \(\text{Edi Ave. and George Street}\)” with speech and a point gesture.

**Types of MMIF**

As shown in Figure 1, in general, multimodal input fusion can occur at two levels: