Abstract

Human communication and interaction are inherently multimodal. Speech, prosody, gesture, gaze, facial displays, and body posture are involved in the communicative act. Hence, understanding witting or unwitting visual displays is critical to both the analysis and synthesis of human communicative behaviour.

One path from multimodal behaviour to language is bridged by the underlying mental imagery. This visuospatial imagery, for a speaker, relates not to the elements of syntax, but to the units of thought that drive the expression (vocal utterance and visible display). The basic idea is that mental imagery is integral to language production, and non-verbal behaviour (gesture, gaze, facial expression) informs us of this imagery. This gives us a handle to extract information on human discourse from a multimedia record of multimodal language performance. The question becomes what computable features of behaviour are informative about imagery and the organization of the discourse.

We shall see a feature-decomposition approach that facilitates cross-modal analysis at the level of discourse planning and conceptualization. We shall discuss our experimental framework and cite concrete examples of features that provide computational access to discourse structure. We shall also see several examples of how these principles may be applied in such domains as meeting analysis and distance tutoring.