Semantic-level multimodal integration of video and speech data streams

Olga Vybornova
Universite Catholique de Louvain

Hildeberto Mendonça
Universite Catholique de Louvain

Benoit Macq
Universite Catholique de Louvain

Abstract: A method of multimodal multi-level fusion integrating contextual information obtained from spoken input and visual scene analysis is defined and developed in this research. The resulting experimental application poses a few major challenges: the system has to process unrestricted natural language, free human behaviour, and to manage data from two persons in their home/office environment. The application employs various integrated components: speech recognition, person identification, syntactic parsing, NL semantic analysis, image processing, human behavior analysis, the domain ontology and the decision-making engine.

In our approach the high level fusion of input streams is performed in three stages: (i) early fusion that merges information at the recognition stage to provide reliable reference resolution, (ii) late fusion that integrates all the information at the final stage to interpret the human behavior, and (iii) reinforcement fusion that executes one or more cycles of verification, reevaluating all identified meanings to improve the overall integration.