Social interaction is a fundamental aspect of human life and is also a key research area in psychology and cognitive science. Social psychologists have been researching the dimensions of social interaction for decades and found out that a variety of social communicative cues strongly determine social behavior and interaction outcomes. Many of these cues are consciously produced, in the form of spoken language. However, besides the spoken words, human interaction also involves nonverbal elements, which are extensively and often unconsciously used in human communication. Nonverbal communication is conveyed as [wordless](http://en.wikipedia.org/wiki/Word) messages, in parallel to the spoken words, through aural cues (voice quality, speaking style, rhythm, intonation) and also through visual cues ([gesture](http://en.wikipedia.org/wiki/Gesture)s; [body language](http://en.wikipedia.org/wiki/Body_language); [facial expression](http://en.wikipedia.org/wiki/Facial_expression) and [gaze](http://en.wikipedia.org/wiki/Eye_contact)). These nonverbal cues are used by all of us every day to infer the mood and personality of others, as well as to make sense of social relations, in a very wide range of situations.

Computational analysis of social interaction, in particular of face-to-face group conversations is an emerging field of research in several communities such as human computer interaction, machine learning, speech and language processing, and computer vision. Close connection with other disciplines including psychology and linguistics also exist in order to understand what kind of verbal and non-verbal signals are used in diverse social situations to infer human behavior. The ultimate aim is to develop computational systems that can automatically infer human behavior by observing a group conversation via sensing devices such as cameras and microphones. Besides the value for several social sciences, these systems could open doors to a number of relevant applications that support interaction and communication, including tools that improve collective decision making, that help keep remote users in the loop in teleconferencing systems, and that support self-assessment, training, and education.

Our aim in the “Automatic Analysis of Group Conversations via Visual Cues in Nonverbal Communication (NOVICOM)” project, is to develop computational systems that can automatically analyze social behavior by observing conversations via cameras and microphones. We focus on group conversations and aim to infer aspects of the underlying social context, including both individual actions and interactions among the people in the group. Examples to such aspects are dominance, leadership, and roles.

**The NOVICOM project**

In the NOVICOM project, conducted at the Social Computing group at Idiap, we are exploring models that can estimate social behavior from both audio and visual nonverbal cues, with a specific focus on visual cues. We concentrated on a selected number of key research tasks in social interaction analysis. These include the automatic estimation of dominance in groups, the emergence of leadership, and personality. In these situations, people unconsciously display visual cues, in the form of gestures and body postures, which partly reveal their social attributes. For each task, our specific objectives are twofold. First we attempt to automatically detect the visual nonverbal cues that are displayed during interaction. Second, we investigate multimodal approaches that integrate audio and visual nonverbal cues to infer social concepts.

For our research, we used publicly available data emerging from previous EC-funded projects at Idiap, as well as newly collected data. For the new data, we were mainly interested in recording natural conversations. For this purpose, we have designed a portable audio-visual recording system, which includes two webcams and a microphone array and is able to record four people sitting around a table. This work has involved in an international collaboration with a startup company that is developing new audio capture devices. Our system allows for the capture of group interaction outside the laboratory with people who volunteer to participate in a number of group discussion scenarios.

As a first task, we have concentrated on the modeling of dominance. We have investigated the effect of visual nonverbal cues to estimate the most dominant and least dominant person in a group conversation. For each meeting, we process the audio and visual recordings obtained from microphones and cameras. For the audio, one crucial task is to segment each speaker (i.e. who speaks when) from the microphone data. For the video, we find where the participants are and track their visual activity. Once we have this information, we extract several audio-visual nonverbal features for each participant, including audio features like speaker turns, interruptions, and speech energy; and visual features such as total body activity, usage of head gestures, and hand gestures. For inference, we developed multimodal fusion techniques to utilize audio and visual nonverbal information jointly. We are able to achieve accuracies around 90% to estimate the most dominant person in the meeting, in comparison to human judgments, using fused audio and visual nonverbal cues. Another task we have worked on is the identification of emergent leadership in small groups. For this purpose, we collected a new dataset with the portable recording system, which includes meeting recordings of newly formed groups trying to solve a task. We performed a multimodal fusion approach to fuse different audio and visual nonverbal cues and we observed around 80% accuracy of estimating the emergent leader in the group. Both of these studies showed that the visual information is necessary and should be used together with audio in order to achieve better performance in dominance and emergent leadership estimation. As another dimension of social behaviour analysis, we investigated the use of automatically extracted audio and visual nonverbal cues as descriptors of personality. We showed that the combination of audio and visual cues can be used to predict 34% of the variance for the Extraversion personality trait, which corresponds to the most reliably judged personality trait.

Last but not least, there is the issue of where and how to apply these new technologies. In the last few years there is a strong interest on analyzing human activity, especially in smart room applications. Among the range of applications that provide support systems in rooms equipped with sensing devices, the computational analysis of social interaction is an emerging field of research. Projects like NOVICOM are just the first steps towards analyzing social context in group communication. In their current form, most existing technologies are far from performing a complete analysis to be used as a tool in our daily life, but can be seen as modules that are becoming robust very rapidly. The research performed in NOVICOM will ultimately enable the development of techniques that analyze social interactions via visual cues, in addition to the aural ones. These techniques can be used to develop systems that support collective decision making in meetings by providing key information about the meeting, and about the participants behavior. The outcome of the analysis can be used to resolve conflicts, creating attention and setting the focus, where appropriate. Alternatively in a teleconferencing system, these techniques can help to keep the interest level of the remote users. They can also be used to develop tools for self assessment, training, and education. Thus, when an accurate enough automatic analysis of group interactions is achieved through this research, the benefit gained will be the development of tools that can be used to support and enhance our work and our daily life.

**Contact Information:**

NOVICOM website : <http://www.idiap.ch/project/novicom/>

Oya Aran, Marie Curie fellow : oya.aran@idiap.ch

Daniel Gatica-Perez, Scientist in Charge: gatica@idiap.ch