Social Computer Vision for Group Behavior Analysis

Marco Cristani

Università di Verona, Verona, Italy. marco.cristani@univr.it

Detecting human interactions represents one of the most intriguing frontiers of the automated surveillance since more than a decade. Recently, sociologic and psychological findings have been considered into video surveillance algorithms, especially thanks to the advent of Social Signal Processing work, a recent multi-disciplinary area where computer vision and social sciences converge.

This chapter follows this direction and proposes a detailed overview of the work we are conducting to detect and characterize social interactions, with particular reference to groups. In particular, we will present three scenarios where a group of interacting people is first detected, using positional and orientation features, and then characterized by extracting voice activities from solely visual cues. Finally, we classify existent social relations among the participants.

The first contribution is devoted to detect social interactions using statistical analysis of spatial-orientation arrangements that have a sociological relevance. As social interactions we intend the acts, actions, or practices of two or more people mutually oriented towards each other. In general, any dynamic sequence of social actions between individuals (or groups) that modify their actions and reactions by their interaction partner(s).

We analyze quasi-stationary people in an unconstrained scenario identifying those subjects engaged in a face-to-face interaction, i.e., a scene monitored by a single camera where a variable amount of people (10-20) is present. We import into the analysis the sociological notion of F-formation as defined by Adam Kendon in the late ’70s [1]. Simply speaking, F-formations are spatial patterns maintained during social interactions by two or more people. Quoting Kendon, *“an F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access.”*.

In practice, an F-formation is a set of possible geometrical configurations in space that people may assume while participating in a social interaction There can be different F-formations: in the case of two participants, typical F-formation arrangements are vis-a-vis, L-shape, and side-by-side; when there are more than three participants, a circular formation is typically formed. Our approach aims at automatically detecting these visually significant configurations taking as input a calibrated scenario, in which the position of the people and their head’s orientations have been estimated. In particular, we design an F-formation recognizer based on a Hough-voting strategy, which lies between an implicit shape model, where weighted local features vote for a location in the image plane, and a mere generalized Hough procedure where the local features have not to be in a fixed number as in the implicit shape model. This approach provides the estimation of the position of the social space subject to the interaction, so as the persons involved, thus individuating the people which are socially interacting.

In such regard, our approach is the first to use F-formations detection in order to discover social interactions solely from visual cues. It has been tested on about a hundred of simulated scenarios, and two real annotated datasets. In the latter two cases, tens of individuals are captured while they were enjoying coffee breaks, in indoor and outdoor environment, giving rise to heterogeneous real crowded scenarios. Our method obtains convincing results and it compares favorably against other similar work aimed at group detection.

Following the analysis of non-verbal cues for the detection of social signals, our second contribution is related to the characterization of the group interactions by proposing a Voice Activity Detection approach only based on the automatic measurement of the persons’ gesturing activities.

This work takes inspiration from the fact that it is common experience to observe that people accompany speech with gestures, the *"... range of visible bodily actions that are, more or less, generally regarded as part of a person`s willing expression ..."* [2]. Far from being independent phenomena, speech and gestures are so tightly intertwined that every important investigation of language has taken gestures into account, from *De Oratore* by Cicero (1st Century B.C.) to the latest studies in cognitive sciences [3,4] showing that the two modalities are *"... components of a single overall plan ..."* [2].

Hence, this work presents a method for estimating the level of gesturing as a means to perform Voice Activity Detection (VAD), i.e. to automatically recognize whether a person is speaking or not. The main rationale is that audio, the most natural and reliable channel when it comes to VAD, might be unavailable for technical, legal, privacy related issues or simply for a noisy scenario. A condition that applies in particular to surveillance scenarios where people are monitored in public spaces and are not necessarily aware of being recorded.

Previous works take advantage of restrictive experimental setups, such as a meeting room where the audio and video signals can be extracted with a strong degree of reliability. This allows to perform an accurate diarization of the audio signal (e.g., [Ba2010]) in a smart meeting room, which remarkably simplifies the VAD by using multiple cameras capturing each person individually at close distance. Our work breaks such constraints and deploys a system "in the wild" designing a more credible setup for a video surveillance system. We use solely visual cues obtained from only one camera positioned 7 meters above the scene. In particular, the experiments focus on people involved in standing conversations, with an automatic person tracking system that follows each individual. Our VAD method is based on a local video descriptor that extracts from each individual body the optical flow, encoding its energy and "complexity" using an entropy-like measure. This allows one to discriminate between body oscillations or noise introduced by the tracker, where the optical flow is low and homogeneous, and genuine gestures, where the movement of head, arms and trunk produces a local flow field which is diverse in both intensity and direction. The descriptor extracted for each participant produces a signal that can be used for VAD.

The proposed approach is interesting for three main aspects. First, the relationship between speech and gestures has been widely documented and studied, but relatively few quantitative investigations of this phenomenon have been made. Second, approaches similar to ours might help to infer information about privacy protected data (speech in this case) from publicly accessible data (gestures in this case): this is also important for establishing whether the simple absence of a certain channel is sufficient to protect the privacy of people and how much. Finally, inferring missing data from available ones can make techniques dealing with challenging scenarios more effective and reliable.

The results show that our gesturing level estimation approach proposed achieves, on a frame-by-frame basis, an accuracy of 71% in distinguishing between speech and non-speech for the tested cases.

The last contribution is about the characterization of the group interaction aimed at the recognition of the relations among the interlocutors by using proxemic cues.

Proxemics can be defined as "[...] the study of man's transactions as he perceives and uses intimate, personal, social and public space in various settings [...]", quoting Hall [1], [2], the anthropologist who first introduced this term in 1966. In other words, proxemics investigates how people use and organize the space they share with others to communicate, typically outside conscious awareness, socially relevant information such as personality traits (e.g., dominant people tend to use more space than others in shared environments), attitudes (e.g., people that discuss tend to seat in front of the other, whereas people that collaborate tend to seat side by side), etc..

This work focuses on one of the most important aspects of proxemics, namely the relationship between physical and social distance. In particular, the paper shows that interpersonal distance (measured automatically using computer vision techniques) provides physical evidence of the social distance between two individuals, i.e. of whether they are simply acquainted, friends, or involved in a romantic relationship. The proposed approach consists of two main stages: the first is the automatic measurement of interpersonal distances, the second is the automatic analysis of interpersonal distances in terms of proxemics and social relations.

The choice of distance as a social relation cue relies on one of the most basic and fundamental findings of proxemics. In fact, people tend to unconsciously organize the space around them in concentric zones corresponding to different degrees of intimacy [5], [6]. The size of the zones changes with a number of factors (culture, gender, physical constraints, etc.), but the resulting effect remains the same: the more two people are intimate, the closer they get. Furthermore, intimacy appears to correlate with distance more than with other important proxemic cues like, e.g., mutual orientation. Hence, it is reasonable to expect that the distance accounts for the social relation between two persons.

One of the main contributions of the paper is that the experiments consider an ecological scenario (standing conversations) where more than two people are involved. This represents a problem because in this case distances are not only determined by the degree of intimacy, but also by the need of ensuring that every person can participate to the interaction. This leads to the emergence of stable spatial arrangements, the already mentioned F-formations, that impose a constraint on interpersonal distances and need to be detected automatically solely using unobtrusive computer vision techniques. Furthermore, not all distances can be used because, in some cases, they are no longer determined by the degree of intimacy, but rather by geometric constraints.

Our approach is to consider only the distances between people adjacent in the F-formation and, unlike other works in the literature, the radii of the concentric zones corresponding to different degrees of intimacy are not imposed a-priori, but rather they are learned from the data using an unsupervised approach. This makes the technique robust with respect to the factors affecting proxemic behavior, like culture, gender, etc., as well as environmental boundaries. In particular, the experiments show how the organization into zones changes when decreasing the space available to the subjects and how the unsupervised approach is robust to such effect.

Standing conversations are an ideal scenario not only because they offer excellent examples of proxemic behavior, but also because they allow one to work at the crossroad between surveillance technologies, often applied to monitor the behavior of people in public spaces, and domains like Social Signal Processing [7] that focus on automatic understanding of social behavior. All the presented techniques draws a path towards a new paradigm for analyzing social interactions where social signals can be extracted in unconstrained scenarios. This is expected to lead, on the long-term, to socially intelligent surveillance and monitoring technologies [8].

References

1. Kendon, A. (1990). *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge University Press.
2. Kendon, A. (2000). Language and gesture: unity or duality? In *Language and Gestures*, D. McNeill Eds., Cambridge University Press, 47-63.
3. McNeill, D.(1992). *Hand and mind: What gestures reveal about thought*. Chicago University Press.
4. Kendon, A. (1980). *Gesticulation and speech: Two aspects of the process of utterance*. In: The Relationship of verbal and nonverbal communication, M.R.Key Eds., Mouton Publishers, The Hague, 207-227.
5. Hall, E.T. (1974). *Handbook for proxemic research. Studies in the anthropology of visual communication series. Society for the Anthropology of Visual Communication*.
6. Hall, E.T. (1966). *The hidden dimension*. Doubleday Eds. New York.
7. Vinciarelli, A., Pantic, M., Bourlard, H. (2009). Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing* **27**, 1743-1759.
8. Cristani, M., Murino, V., Vinciarelli A. (2010). Socially intelligent surveillance and monitoring: analysing social dimensions of physical space, *Int'l Wks on Socially Intelligent Surveillance and Monitoring*, 51-58.