Socially Intelligent Surveillance and Monitoring: Analysing Social Dimensions of Physical Space

M. Cristani[1,2], V. Murino[1,2], A. Vinciarelli[3,4]
[1]*Computer Science Department, University of Verona,* Italy
[2]*Istituto Italiano di Tecnologia (IIT),* Genova, Italy
[3]*University of Glasgow,* UK
[4]*Idiap Research Institute,* Martigny, CH

Abstract

In general terms, surveillance and monitoring technologies aim at understanding what people do in a given environment, whether this means to ensure the safety of workers on the factory floor, to detect crimes occurring in indoor or outdoor settings, or to monitor the flow of large crowds through public spaces. However, surveillance and monitoring technologies rarely consider that they analyze human behavior, a phenomenon subject to principles and laws rigorous enough to produce stable and predictable patterns corresponding to social, affective, and psychological phenomena. On the other hand, these phenomena are the subject of other computing domains, in particular Social Signal Processing and Affective Computing, that typically neglect scenarios relevant to surveillance and monitoring technologies, especially when it comes to social and affective dimensions of space in human activities. The goal of this paper is to show that the investigation of the overlapping area between surveillance and monitoring on one side, and Social Signal Processing and Affective Computing on the other side can bring significant progress in both domains and open a number of interesting research perspectives.

1. Introduction

Like anthropologists say it, "*space speaks*". Whenever left free to move in a large environment (e.g., the hall of a hotel, a square, a street, a garden, a waiting room, a restaurant, etc.), people seem to wander without precise criteria, but actually respect patterns and trajectories largely dominated by social mechanisms. An invisible bubble seems to surround each person and keeps people far from one another unless there are physical constraints or the space is too crowded [33]. Social bonds, whether they involve only two persons or a large group of individuals, are shown by decreasing mutual distances between people and by delimiting regions of space close to others [38]. Social messages like dominance, inclusion, exclusion and rapport are communicated via mutual positioning, body orientation and posture [39, 58, 61].

In principle, all above phenomena are accessible to automatic analysis through computer vision and pattern recognition, the main domains used for automatic surveillance so far [31, 63]. Observation activities have never been as extensive as today and they keep increasing in terms of both amount and scope. Furthermore, involved technologies progress at a significant pace (some sensors exceed now human capabilities) and, as they are cheap and easily available on the market, have an increasingly large diffusion. This does not happen by chance: automatization makes observation objective and rigorous while safer, personnel does not need to be present in a potentially dangerous environment, and more extensive, public and private ambients can be monitored 24 hours a day from several points of view with limited human intervention.

One of the main challenges for a surveillance system is the automatic recognition of atypical (i.e., dangerous or suspicious) behaviors in video recordings. This is usually accomplished using a serial architecture built upon an array of techniques aimed at extracting low-level information. This includes, for example, foreground/background segmentation [56, 57, 7] and object tracking [24]. After these early steps, high-level analysis approaches detect atomic actions (e.g., gestures) as well as complex activities (i.e., spatio-temporal structures composed of atomic actions) [12], possibly exploiting ontologies for ensuring interoperability across different platforms and semantic descriptions understandable to human operators [23]. Typical application cases are traffic monitoring and surveillance of large areas (e.g., courts or parks) and meetings, with a preference for relatively stable and predictable scenarios.

However, these technologies seem to forget that, for human beings, physical and social space are tightly intertwined and no intelligent monitoring is possible without taking into account social aspects associated to behaviors displayed under the eyes of the cameras. This is especially regrettable when other domains, e.g. Affective Computing (AC) [55] or Social Signal Processing (SSP) [66], pay significant attention to social, affective and emotional aspects of human behavior, but do not take into account scenarios relevant to surveillance and monitoring, especially when it comes to the analysis of large groups of people sharing the same physical space.

The goal of this paper is to show how the application of socially and emotionally aware approaches (like those developed in SSP and AC, respectively) promises to improve the state of the art in surveillance and monitoring, while the use of surveillance and monitoring technologies in SSP and AC can allow the automatic analysis of social and affective phenomena investigated so far only in human sciences (e.g., proxemics and territoriality).

In the following, Section 2 shows that the analysis of social phenomena related to the use of physical space has been, at least so far, neglected by socially and emotionally aware technologies. Section 3 presents a state of the art in surveillance and monitoring showing that social aspects of typical application scenarios are almost never taken into account. Section 4 outlines the most important research questions arising from the cross-pollination between the two domains, as well as the applications most likely to profit from the inclusion of socially-aware technologies in surveillance and monitoring (and viceversa). Finally, Section 5 draws some conclusions.

2. Socially and Emotionally Aware Technologies

Socially and emotionally aware technologies like Social Signal Processing and Affective Computing aim at bridging the social and emotional intelligence gap, respectively, between humans and machines (see [66] and [72] for extensive surveys). In the case of both SSP and AC, the core aspect is the automatic analysis of nonverbal behavior in face-to-face interactions with the goal of inferring information such as mutual attitude between interactants, roles, conflict, personality traits, dominance, emotions, etc. [68]. The focus on nonverbal communication comes from several decades of investigation in human sciences (psychology, anthropology, sociology, etc.) showing that humans use nonverbal behavioral cues like facial expressions, vocalizations (laughter, fillers, back-channel, etc.), gestures or postures to convey, often outside conscious awareness, *social signals*, i.e. their attitude towards interactions and social environments, as well as emotions [39, 58].

The literature has identified a large number of behavioral cues carrying social meaning. These have been grouped into five classes called *codes* [30]: *Physical Appearance* (attractiveness, clothes, ornaments, somatotype, etc.), *Vocal Behavior* (everything else than words in speech), *Face and Eyes Behavior* (expressions, gaze, head pose, etc.), *Gestures and Postures* (bodily movements, conscious and unconscious gestures, orientation with respect to others, etc.), *Space and Environment* (mutual distances, spatial organization of people, territoriality, etc.).

Computer science has developed approaches for the automatic analysis of many aspects of the first four codes (with the exception of appearance that has been addressed in a relatively few works), but no major efforts have been done, to the best of our knowledge, towards the automatic inference of socially and emotionally relevant information from the way people use, organize and share their physical space [66, 67]. In contrast, this subject has been extensively investigated in human sciences where the spatial arrangement of people in social encounters has been shown to be a reliable evidence of the social phenomena taking place among interacting individuals [28, 32, 33, 38].

There are two main reasons behind the lack of attention towards automatic analysis of the way people use space. The first is that computing domains aimed at understanding social and affective behavior have focused on face-to-face small group interactions [26, 66]. These have been the subject of major attention because, one one hand, they represent the most common and primordial form of human-human interaction and, on the other hand, they involve those social and emotional phenomena that most affect our life like conflict, exclusion, affiliation, roles, dominance, personality, performance, etc. [41].

The second important reason is that the analysis of behavioral cues like facial expressions, prosody, small gestures, etc. requires to perform experiments in a controlled setting like a smart meeting room or a laboratory. This is incompatible with surveillance and monitoring scenarios that must take place in natural, non-constrained settings to be sufficiently realistic. In other words, it is difficult to imagine a scenario where both the analysis of spatial behavior and the analysis of finer behavioral cues is possible. However, this corresponds exactly to our way of perceiving the world, where subtler behavioral cues become important only below a certain distance ($1-2$ meters), while they do not play a major role at larger distances [32, 33].

Thus, the application of automatic analysis approaches to the spatial organization of social encounters and, more in general, to the social and emotional dimensions of space, not only fills a gap in the state of the art of SSP and other disciplines aimed at understanding social interactions, but also represents a research opportunity alternative to any scenario considered so far in socially aware technologies.

3. Surveillance and Monitoring

In the evolution of video surveillance systems, the very first generation was formed by motion detectors, detecting any movement in the camera view [43]. Even if the advantage with respect to the human capabilities was consistent (in terms of endurance), those systems were affected by a dramatic false alarm rate. This because all the motion was assumed as synonym of threat.

The next step of the evolution embedded the detection and the tracking of "objects" of interest in a serial multistage framework [34, 70]. The lower levels are those closer to the raw data, performing operations such as noise removal, and motion detection. The latter is typically carried out initializing and updating a dynamic background model [56], and pixels or pixels' areas deviating from the background model statistics are labeled as foreground, i.e., an entity worth to be further analyzed (see Fig. 1 b).

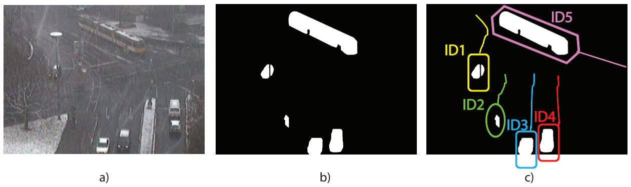


Figure 1. The second generation of surveillance systems: multilayer architectures. a) Input frame; b) background subtraction; c) detection and tracking of different entities.

Higher levels carry out detections and tracking of "objects" on the foreground pixels, thus reducing the false alarm rate and giving more flexibility to the user in terms of defining entities of interest (see Fig. 1c).
The third evolution step added a further higher level to the previous architecture, devoted to reasoning about the trajectories of the foreground objects [52, 29, 62]. This level allows the user to specify events of interest as particular trajectories, drawn by a precise class of objects (see Fig. 1c). In this case, it becomes feasible to model and retrieve events like "pedestrian crosses the road on the crosswalk", considering a priori manual definition of a set of pre-determined normal and abnormal events. In another spirit, similar systems encapsulate the capability of learning in an unsupervised way what is usual in a given scenario and what is not, considering sufficient statistics of trajectories [29, 62].
In the meanwhile, multi-camera and multi-object tracking methods are becoming able to track multiple objects across far locations, captured by sensors with non overlapping camera views [64, 65]. These systems face several objects at-a-time (around 10-15), succeeding to manage important problems such as occlusions (an object disappears and reappears in the same camera field of view), and re-identification (an object disappears and reappears across different non overlapped camera views) (see Fig. 2) [5, 73, 20].

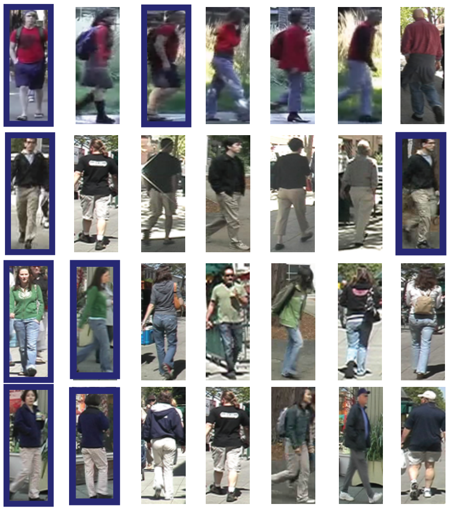


Figure 2. The re-identification problem: once a specific person has been detected, he/she has to be followed across different times and locations. On each row, two views of the same individual (highlighted in blue) are shown, whereas the others are similar subjects.

For a deeper and more detailed taxonomy of video surveillance systems, see [37]. Summarizing, the last generation of surveillance systems witnesses a certain maturity in managing the lower levels of the data processing, i.e, dealing with multiple visual entities, capturing their (even occluded) positions in a given possibly sparse environment. However, considering the highest processing level, much more can be done. In the following, we list different scenarios where the lack of social knowledge clearly emerges.

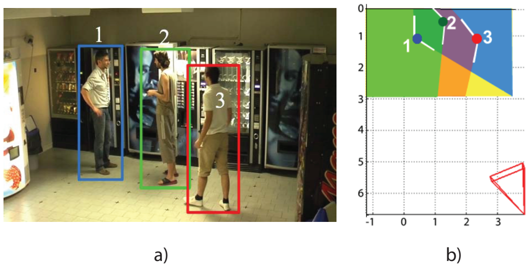
- **Definition of behavior**. In the surveillance literature, many approaches present applications of behavior profiling for activity analysis. Some of them avoid to focus on an operational translation of behavior [11, 70], often employing the notions of behavior and activity interchangeably. Other approaches often propose ad-hoc definitions, well-suited for the task at hand [8, 54, 34, 2, 9, 17, 59, 35, 18], generating a bunch of diverse characterizations. For example, in the early analysis proposed in [8], the behavior is a hierarchical entity tightly connected with the notion of human motion or *action*; More specifically, a *movement* is the most basic brick, requiring no contextual or sequential knowledge to be recognized; the *action* is a larger scale event formed by ordered movements, which typically includes interaction with

the environment and causal relationships. Similar structural definitions can be found in several works [34, 2, 9, 59, 35, 18]. In [54], human behaviors is accurately described as a set of dynamic models (e.g., Kalman filters) sequenced together by using a Markov chain. Markov dynamics is typically a common choice for characterizing the time evolution of a behaviour. In [25, 35], time duration of actions is explicitly taken into account employing a Markov logic approach. A hierarchical notion of behavior, that enriches the serial action-based definition by adding different abstraction levels, is proposed in [18, 49].
In [17], a completely novel concept of behavior is proposed, which consists of an ensemble of spatio-temporal feature points, deeply investigated in [50]. However, these structured definitions are based on the design of visual bricks (visual words) that do not carry any intuitive meanings.

From one side, all these definitions witness the multi-faceted nature of behavior as an intuitive concept, but, from the other side, they also highlight the lack of a common (and general) notion of behavior. Just to tackle this problem, SSP can provide important suggestions aimed at finding an accurate structural definition of behavior or a common ground for reasoning on behaviors, widely accepted and used by the several scientific communities.

- **Definition of threatening behavior.** In almost every surveillance system published in literature, the main goal is that of promptly identifying threatening behaviours in an automatic way. Assuming that the meaning of behavior has been grounded, the underlying, subtle, problem is that the meaning of ”threatening” is actually unspecified, and reduced (too) often to that of “abnormal” or “unexpected” [62, 18]. This translates in having complex techniques that simply collect a statistics of trajectories, and whenever a different trajectory does hold, it is labelled as abnormal and considered as potentially threatening. Especially when the statistics collected is scarce, this will cause huge amounts of false positives, making the system unusable for practical purposes. SSP may help in giving a priori knowledge on what is really threatening, and what is simply a reasonable deviation from a common behavior. Moreover, it can also be interesting to learn directly, on the basis of examples, what is a dangerous or suspect situation from those that are not, maybe using a semi-supervised approach with the support of a human operator.

- **Definition of group.** In the recent surveillance approaches, tracking applications are undoubtedly the workhorses, focusing on each person in the scene, capturing its trajectory, helping in analyzing its motion, gestures, etc.. Recently, the focus has been moved beyond the mere multi-object tracking, considering the groups as interesting entities [27, 47, 16, 69, 53, 42, 44, 22, 40]. Capturing groups of people helps in defining a visual context where a particular person may be better recognized [73], it enriches the expressivity of a surveillance profiling, and is indeed necessary when the number of people in the scene is too high for employing the simultaneus tracking of multiple persons. But what is a group? Usually this corresponds to having a set of individuals exhibiting similar characteristic, i.e., close in space, with the same oriented motion and similar velocity, possibly wearing similar dresses. This description fails in effectively performing a high-level semantic surveillance profiling: for example, it fails in distinguishing a situation where space constraints force the people to stay close from that where the proximity is the result of a common intention of several subjects. Social Signal Processing may help in these cases, providing novel cues that can be exploited by standard surveillance algorithms for identifying groups by a co-existence of social bonds among individuals. For instance, in a little dense social situation, the face orientation of each person may help to realize the persons who are known to each other (i.e., identifying a group) [21] and who are not (see Fig. 3).



a) b)

Figure 3. Group detection under a social signalling perspective: a group is composed by a set of people, that are 1) close and 2) looking at each other [21]. Such properties can be extracted from the videosequence using calibrated tracking and head pose estimation techniques. Above, a) an instance of tracking; b) the projection of the people tracked on the ground plane, with related fields of view portrayed as coloured, semi transparent areas (blue for object 1, green for object 2 and red for object 3). Subject 1 and 2 are forming a group.

- **Definition and detection of interaction.** In the literature, the modelling of human interactions has reached a deep maturity, both under a pure visual perspective [71, 51], and also considering other modalities [60, 4, 46, 15, 45, 14, 10]. Anyway, like in the case of the definition of behavior, there is the lack of a common, formal definition of interaction. Having a universal notion of interaction may lead to very interesting novel applications. For example, a topic that has not received attention so far is that of the *detection* of interactions. In other words, all the above quoted studies face the problem of the interaction modelling in very constrained scenarios (meetings, games, etc.), where interacting activities are foreseen or expected. None of them takes explicitly into account the problem of detecting an interaction in a scenario where interactions may appear or not.

More in detail, what is ultimately needed is a definition of interaction between people in its several forms: individual vs individual, individual vs group and vice versa, group vs group, in a crowd. The kinds of interactions range from a simple face-to-face discussion to more subtle signals, not necessarily in terms of speech, often in terms of body motion, face expression (indicating a sort of feeling or emphaty), or pre-defined hints that can be exchanged by the involved persons, clues that can be conscious or not. In the SSP and particularly in the social science literature, there is much material on these domains, but what is missing is a clear categorization and identification of the several types of interactions in the diverse contexts, and, especially, it is unclear if computational techniques can be able to detect and classify such interactions and to what extent, possibly predicting their final objectives.

Another very recent kind of video surveillance system, not falling within the above taxonomy, is that of the surveillance of crowds. In this case, the challenge for the video surveillance community is to move from the typical scenario where about 20 subjects may act, to that of a huge mass of humans (100-200+ individuals) moving on a wide area [48, 36, 1]. In this case, the state of the art of the research is unripe, and there are no clues on what to look for. Usually, the idea is to model the motion flow of the mass of people, individuating if the flow is always the same during a certain period, or if it changes suddenly. In this case, SSP may help in exploiting standard sociologic foundations, and telling what could be intended as dangerous or not in a similar scenario.

In summary, none of the techniques present so far in the state of the art is using SSP and AC findings or hints, only very recently this need is becoming explicit [21, 19].

4. Applications and Research Questions

The previous two sections have shown that, on one hand, SSP and AC have neglected surveillance relevant scenarios and the use of space as a source of information about social and affective phenomena and, on the other hand, that surveillance technologies do not take into account social, affective and emotional aspects of human behavior, even if they are expected to analyze what people do and how they interact in a given environment.

The cross-pollination between surveillance and socially- and emotionally- aware technologies opens several research questions and can lay the ground for several applications. The most important research questions we foresee are listed in the following.

- *Is it possible to infer social phenomena from the spatial organization of people in public spaces following the findings of human sciences?* Psychologists and anthropologists have clearly shown that people use space, mutual orientation, and distances to communicate socially relevant information such as quality of rapport, sustained interaction, inclusion and exclusion, etc., but it is not clear whether an automatic analysis of such phenomena is possible, especially in real-life situations. Actually, in this cases, non optimal location of the sensors, the complexity of the scene (e.g., due to the high number of people involved), and the difficulty of capturing the subtle signals exchanged by the persons, may affect the success of this analysis.

- *Is it possible to apply human sciences' findings to identify threatening, criminal or other surveillance relevant behaviors?* Most surveillance approaches can identify unusual behaviors, but cannot assess how dangerous they are. Socially and emotionally aware technologies hold the promise of improving the interpretation of observed behaviors, but extensive experiments are needed to show that it is actually the case. In this context, one issue to cope with lies on the fact that simulated data (lab experimentations) are likely not capturing the complexity of a real behavior, and, on the other hand, it is not easy to have the availability of real behaviors, which needs to be annotated as well.

- *Is it possible to influence social behavior of people by modifying the physical setup of their interactions?* Social interactions are well known to be influenced by the physical setup of the place where they take place. However, the setup is typically defined manually and no attempts have been done to do it automatically, possibly in reaction to phenomena actually taking place at a given moment.

- *Is it possible to better understand social interactions through automatic analysis of spatial behavior?* Human sciences rely more and more on automatic approaches to identify principles and laws underlying social interactions. Thus, the automatic analysis of spatial behavior can help human sciences to better understand dynamics and laws of social behavior.

Addressing the above questions is likely to foster new applications as well or, at least, to improve existing applications like those listed in the following.

- *Design of public spaces*. Simple architectural elements are known to influence significantly the collective behavior of large crowds in public spaces [3]. Socially intelligent surveillance technologies can help to analyze this phenomenon and improve the design of public spaces like train stations, airports, squares, etc. that are typically populated by large amounts of interacting individuals.
- *Marketing*. Consumer behavior in retail spaces is affected by several situational variables including the physical setup of the shops [6]. Automatic behavior analysis can be an important tool for marketing analysis aimed at understanding what are the products attracting more attention, what are the main obstacles towards effective customer-seller interactions, what is the best position for a product, etc..
- *Learning spaces*. The effectiveness of a learning space is heavily influenced by its physical setup, especially when the learning process requires the collaboration of many individuals [13]. Socially and emotionally intelligent surveillance technologies can help the design of effective learning environments by understanding those behavioral processes that help or compound effective collaboration between people.

Application domains and research questions proposed in this section are certainly not an exhaustive list, and many other possibilities are available for jointly applying surveillance and socially aware technologies.

5. Conclusions

This paper has pointed out that socially-aware technologies tend to neglect surveillance scenarios to focus on face-to-face small group interactions, while surveillance and monitoring technologies tend to neglect social, affective and emotional aspects of human behavior even if this is, in ultimate analysis, their main subject of interest. Furthermore, the article has shown how mutual cross-pollination between these two kinds of technologies could lead to new research questions as well as to application domains that, so far, have not been the subject of attention in the computing community (e.g., the design of public spaces or learning environments).

The advantages of socially intelligent surveillance and monitoring would be evident in the two originating domains as well. On one hand, socially aware approaches can address problems that have been neglected so far (in part for lack of computer vision technical competences) such as territoriality, spatial organization of social encounters, social meaning of physical distances, etc.. On the other hand, surveillance technologies, by including socially aware components, could address open issues such as the distinction between normal and threatening behaviors, the dynamics of large crowds, etc., in essence, providing a structured and systematic basic definition of behavior which could be widely accepted and used in the communities.

In conclusion, we are deeply convinced that the cross-fertilization of human and computer sciences, started with the SSP and AC research domains, is going to be inevitably extended to other fields and applications, like surveillance and monitoring, and only in this way a new generation of surveillance systems can be designed, making the necessary jump to go beyond the current technology, so far advanced in incremental steps. Even if this technology can be perceived like a sort of “big brother”, it is indeed important to be investigated to cope with, possibly prevent, the threats and the criminal (mainly terrorist) actions we see too much often around the world. Indeed, this novel technology may also be useful and productive to understand the real bases of human interactions, supporting human and social scientists to go more deeply to the foundations of the social cohabitaion, ultimately contributing to live in a better world.

References

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR07*, pages 1–6, 2007.

[2] J. B. Arie, Z. Wang, P. Pandit, and S. Rajaram. Human activity recognition using multidimensional indexing. *PAMI*, 24(8):1091–1104, August 2002.

[3] P. Ball. *Critical mass: How one thing leads to another*. William Heinemann Ltd, 2004.

[4] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interaction with the influence model. Technical Report 539, MIT MediaLab, 2001.

[5] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded Up Robust Features. In *Proceedings of the European Conference on Computer Vision*, pages 404–417, 2006.

[6] R. Belk. Situational variables and consumer behavior. *Journal of Consumer Research*, 2(3):157–164, 1975.

[7] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger. Review and evaluation of commonly-implemented background subtraction algorithms. In *Proceedings of International Conference on Pattern Recognition*, pages 1–4, 2008.

[8] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. In *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, pages 1257–1265, 1997.

[9] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[10] M. Brand, N. Oliver, and S. Pentland. Coupled hidden markov models for complex action recognition. In *Proc.*

of IEEE Conf. on Computer Vision and Pattern Recognition, 1997.

[11] H. Buxton. Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, 21:125–136, 2003.

[12] R. Chellappa, A. Roy-Chowdhury, and S. Zhou. *Recognition of Humans and Their Activities Using Video*. Morgan and Claypool, 2005.

[13] N. Chism and D. Bickford. *The importance of physical space in creating supportive learning environments*. Jossey-Bass Inc Pub, 2002.

[14] T. Choudhury and S. Basu. Modeling conversational dynamics as a mixed memory markov process. In *Proc. NIPS*, 2004.

[15] M. Cristani, A. Pesarin, C. Drioli, A. Perina, A. Tavano, and V. Murino. Auditory dialog analysis and understanding by generative modelling of interactional dynamics. In *Second IEEE Workshop on CVPR for Human Communicative Behavior Analysis*, Miami, Florida, 2009.

[16] F. Cupillard, F. Brémond, M. Thonnat, I. S. Antipolis, and O. Group. Tracking groups of people for video surveillance. In *University of Kingston (London*, 2001.

[17] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 65–72, 2005.

[18] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *CVPR (1)*, pages 838–845, 2005.

[19] M. Farenzena, L. Bazzani, V. Murino, and M. Cristani. Towards a subject-centered analysis for automated video surveillance. In *15h International Conference on Image Analysis and Processing*, 2009.

[20] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition, 2010. Proceedings. IEEE Conference on*, 2010. in print.

[21] M. Farenzena, A. Tavano, L. Bazzano, D. Tosato, G. Paggetti, G. Menegaz, V. Murino, and M. Cristani. Social interactions by visual focus of attention in a three-dimensional environment. In *Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis (PRAI*HBA)*, 2009.

[22] D. Fehr, R. Sivalingam, V. Morellas, N. Papanikolopoulos, O. Lotfallah, and Y. Park. Counting people in groups. In *International Conference on Advanced Video and Signal Based Surveillance*, pages 152–157, 2009.

[23] A. Francois, R. Nevatia, J. Hobbs, and R. Bolles. Verl: An ontology framework for representing and annotating video events. *IEEE MultiMedia*, 12:76–86, 2005.

[24] L. Fuentes and S. Velastin. People tracking in surveillance applications. *Image and Vision Computing*, 24(11):1165 – 1171, 2006.

[25] A. Galata, N. Jonhson, and D. Hogg. Learning variable-length Markov models of behavior. *Computer Vision and Image Understanding*, 81:398–413, 2001.

[26] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[27] G. Gennari and G. Hager. Probabilistic data association methods in visual tracking of groups. In *CVPR 04*, 2004.

[28] E. Goffman. *Behaviour in public places*. Greenwood Press Reprint, 1963.

[29] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *CVPR 98*, pages 22–29, Washington, DC, USA, 1998. IEEE Computer Society.

[30] L. Guerrero, J. DeVito, and M. Hecht, editors. *The nonverbal communication reader: Classic and contemporary readings*. Waveland Press Inc., 1999.

[31] N. Haering, P. Venetianer, and A. Lipton. The evolution of video surveillance: an overview. *Mach. Vision Appl.*, 19(5-6):279–290, 2008.

[32] E. Hall. *The silent language*. Doubleday, 1959.

[33] R. Hall. *The hidden dimension*. Doubleday, New York, 1966.

[34] I. Haritaoglu, D. Harwood, and L. Davis. W^4: real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000.

[35] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In *IEEE International Conference on Computer Vision*, volume 2, 2003.

[36] M. Hu, S. Ali, and M. Shah. Detecting global motion patterns in complex videos. In *ICPR08*, pages 1–5, 2008.

[37] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics*, 34:334–352, 2004.

[38] A. Kendon. *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge University Press, 1990.

[39] M. Knapp and J. Hall. *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 1972.

[40] Y. Lao and F. Zheng. Tracking a group of highly correlated targets. In *IEEE International Conference on Image Processing*, 2009.

[41] J. Levine and R. Moreland. Small groups. In D. Gilbert and G. Lindzey, editors, *The handbook of social psychology*, volume 2, pages 415–469. Oxford University Press, 1998.

[42] J. S. Marques, P. M. Jorge, A. J. Abrantes, and J. M. Lemos. Tracking groups of pedestrians in video sequences. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, volume 9, pages 101–109, June 2003.

[43] J. Matter. Video motion detection for physical security applications. In *Proc. of the 1990 Winter Meeting of the American Nuclear Society*, pages 1–14, 1990.

[44] T. Mauthner, M. Donoser, and H. Bischof. Robust tracking of spatial related components. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, Dec. 2008.

[45] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3), 2005.

[46] D. McFarland. Respiratory markers of conversational interaction. *J. of Speech, Language, and Hearing Research*, 44(128):43–48, 2001.

[47] S. J. Mckenna, S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld. Tracking groups of people. *Computer Vision and Image Understanding*, 2000.

[48] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR09*, pages 935–942, 2009.

[49] P. Natarajan and R. Nevatia. Hierarchical multi-channel hidden semi markov models. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2562–2567, 2007.

[50] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*, 79(3):299–318, 2008.

[51] N. Oliver, B. Rosario, and A. Pentland. Graphical models for recognising human interactions. In *Advances in Neural Information Processing Systems*, 1998.

[52] T. Olson and F. Brill. Moving object detection and event recognition algorithms for smart cameras. In *Proc. DARPA Image Understanding Workshop*, pages 159–175, 1997.

[53] S. K. Pang, J. Li, and S. Godsill. Models and algorithms for detection and tracking of coordinated groups. In *Symposium of image and Signal Processing and Analisys*, 2007.

[54] A. Pentland and A. Liu. Modeling and prediction of human behavior. *Neural Comput.*, 11(1):229–242, 1999.

[55] R. Picard. *Affective computing*. The MIT Press, 2000.

[56] M. Piccardi. Background subtraction techniques: a review. In *SMC (4)*, pages 3099–3104, 2004.

[57] R. Radke, S. Andra, Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005.

[58] V. Richmond and J. McCroskey. *Nonverbal Behaviors in interpersonal relations*. Allyn and Bacon, 1995.

[59] N. Robertson and I. Reid. A general method for human activity recognition in video. *CVIU*, 103(2-3):232–248, 2006.

[60] B. C. S. Basu, T. Choudhury and A. Pentland. Towards measuring human interactions in conversational settings. In *IEEE Int'l Workshop on Cues in Communication (CUES 2001)*, Hawaii, CA, 2001.

[61] A. Scheflen. *Body Language and Social Order*. Prentice-Hall, Inc., 1973.

[62] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):747–757, 2000.

[63] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, November 2008.

[64] M. Valera and S. Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings - Vision, Image, and Signal Processing*, 152(2):192–204, 2005.

[65] S. Velastin and P. Remagnino. *Intelligent Distributed Video Surveillance Systems*. Institution of Engineering and Technology, 2006.

[66] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[67] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the ACM International Conference on Multimedia*, pages 1061–1070, 2008.

[68] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function, and automatic analysis: a survey. In *Proceedings of the International Conference on Multimodal interfaces*, pages 61–68, 2008.

[69] Y.-D. Wang, J.-K. Wu, A. A. Kassim, and W.-M. Huang. Tracking a variable number of human groups in video using probability hypothesis density. In *ICPR*, 2006.

[70] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[71] A. B. Y. Ivanov, C. Stauffer and W. E. L. Grimson. Video surveillance of interactions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 82–89, 1999.

[72] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[73] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC 2009*, 2009.