Human Behavior Analysis in Video Surveillance: a Social Signal Processing Perspective

Marco Cristani[1,2] R.Raghavendra[1] Alessio Del Bue[1] Vittorio Murino[1,2]

[1]*Istituto Italiano di Tecnologia (IIT), Genova, Italy*
[2]*Dipartimento di Informatica, University of Verona, Italy*

Abstract

The analysis of human activities is one of the most intriguing and important open issues for the automated video surveillance community. Since few years ago, it has been handled following a mere Computer Vision and Pattern Recognition perspective, where an activity corresponded to a temporal sequence of explicit actions (run, stop, sit, walk, etc.). Even under this simplistic assumption, the issue is hard, due to the strong diversity of the people appearance, the number of individuals considered (we may monitor single individuals, groups, crowd), the variability of the environmental conditions (indoor/outdoor, different weather conditions), and the kinds of sensors employed. More recently, the automated surveillance of human activities has been faced considering a new perspective, that brings in notions and principles from the social, affective, and psychological literature, and that is called Social Signal Processing (SSP). SSP employs primarily nonverbal cues, most of them are outside of conscious awareness, like face expressions and gazing, body posture and gestures, vocal characteristics, relative distances in the space and the like. This paper is the first review analyzing this new trend, proposing a structured snapshot of the state of the art and envisaging novel challenges in the surveillance domain where the cross-pollination of Computer Science technologies and Sociology theories may offer valid investigation strategies.

Keywords: Video surveillance, Social signal processing, Activity recognition, Behavior analysis, Human computing.

1. Introduction

Since the 90's, human activity analysis has been one of the most important topics in computer vision, becoming an integral part of many video surveillance systems, but also representing a key application in several other everyday scenarios like workplaces, hospitals, and many others. Analyzing activities involved to date the recognition of motion patterns, and the production of high-level descriptions of actions and interactions among entities of interest. Many surveys on activity analysis have been proposed in the literature: the first example is [1], where techniques for the tracking and the recognition of human motion are reviewed; in [2], methods for the motion of body parts, the tracking of human motion using different camera settings and the recognition of activities are reported. In [3], hand and body tracking strategies are discussed, together with techniques for human activity recognition based on 2D and 3D models. A comprehensive review on vision-based human motion analysis spanning the period 2000–2006 is presented in [4]. In [5], statistical models like Dynamic Bayesian Networks are addressed as one of the most suitable tools for activity recognition. An essay on the different components of a typical video surveillance system, with emphasis on the activity analysis, is reported in [6]. The definition of activity as a complex and coordinated organization of simple actions is exploited in [7]. In the same year, a survey on video surveillance systems has been proposed in [8], also discussing about the different public databases available to validate the algorithms. In the very recent review on activity recognition approaches [9], the different strategies are organized as hierarchical and nonhierarchical, and the last ones are further divided in space-time and sequential methods.

All the above mentioned surveys addressed the modeling of the human activities mainly stressing the technological computer vision aspects. In particular, all of them focus on detecting and recognizing explicit actions, in the sense of gestures performed voluntarily by humans, like running, walking, stopping, seating etc.

Recently, the study on human activities has been re-

vitalized by addressing the so-called *social signals* [10], which are nonverbal cues inspired by the social, affective, and psychological literature [11]. This allows a more principled encoding of how humans act and react to other people and environmental conditions. Social Signal Processing (SSP), also named Social Signaling, represents the scientific field aimed at a systematic, algorithmic and computational analysis of social signals, that is deeply rooted in anthropology and social psychology [12]. More properly, SSP goes beyond the mere human activity modeling, aiming at coding and decoding the human *behavior*. In other words, it focuses to unveil the underlying hidden states that *drive* one to act in a determined way, with particular actions. This challenge is motivated by decades of investigation in human sciences (psychology, anthropology, sociology, etc.) that showed how humans use nonverbal behavioral cues like facial expressions, vocalizations (laughter, fillers, back-channel, etc.), gestures or postures to convey, *often outside conscious awareness*, their attitude towards other people and social environments, as well as emotions [13]. The understanding of these cues is thus paramount in order to understand the social meaning of the activities.

As we will see later, only a minority of works adopted the SSP perspective in a video surveillance setting, but recently (i.e., since 5 years) this trend has rapidly grown. Actually, in surveillance, the main goal is to detect threatening actions as soon as possible: therefore, the possibility of doing this by observing the human behavior as a phenomenon subjected to rigorous principles that produces predictable patterns of activities, turns out to be incredibly important.

The aim of this paper is to review the early years of the social signaling oriented approaches for human behavior analysis in a surveillance context, individuating what are the contact points between surveillance and social signalling, how social signalling may improve the human behavior analysis, envisaging and delineating future perspectives.

The rest of the paper is organized as follows. Section 2 illustrates the processing scheme of a typical video surveillance system. The aim of the section is that of contextualizing which modules of a video surveillance strategy may benefit from the intervention of Social Signaling findings. Section 3 is a short overview of the recent advances in the activity analysis, aimed at defining what is achieved with pure Computer Vision and Pattern Recognition methods. Section 4 is the core of the paper, reviewing the most significant contributions that represent the intersection between video surveillance and SSP. Section 5 addresses the analysis of crowd behavior, that recently has become a well-defined trend in surveillance, discussing the importance of embedding social signals in such studies. Finally, Section 6 draws the conclusions and presents the envisaged future perspectives.

2. A basic video surveillance system overview

A typical surveillance system scheme is composed by two parts: a low-level and a high-level part (see Figure 1). Each part is composed by different stages, explained in the following.

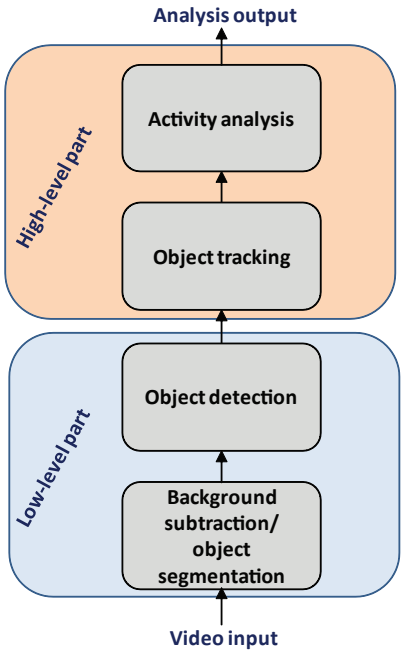

Figure 1: Typical video surveillance automated system.

2.1. The low-level stages

The low-level stages are the background subtraction/object segmentation and the object detection. Such stages preprocess the raw images in order to discover areas of interest.

Background subtraction/object segmentation. Background (BG) subtraction is a fundamental low-level operation that applies on raw videos captured by CCTVs [14]. It aims at learning the expected chromatic aspect of the scene and how it evolves in time, highlighting moving objects (foreground, FG), ideally under a 24/7 policy. Object segmentation follows the background

subtraction and aims at individuating connected regions, pruning away small FG objects, filling holes of large regions, adopting temporal continuity to obtain consistent, smooth regions across time [15].

Object detection. This stage serves to highlight particular classes of targets (humans, vehicles, baggages) in the images. It may be applied on the output of the background subtraction/object segmentation step, or in a dense way over the entire image [16].

These two stage cannot benefit of an intervention of SSP principles, since the processing here is focused on entities, the pixels, carrying very low semantics.

2.2. The high-level stages

The high level stages are the object tracking and the activity analysis.

Object tracking. Tracking is undoubtedly the paramount aspect of any video-surveillance approach, and is very important for the human behavior analysis. For a comprehensive review on tracking for surveillance (out of the scope of this contribution), please read [17]. Tracking aims at computing the trajectory of each distinct object of interest in the scene, associating a ID label and keeping it across occlusions and multiple cameras. A general tracker can be characterized by three main phases: 1) the initialization phase localizes the target that needs to be tracked. It usually relies on heuristic mechanisms combined with some object detector. 2) The dynamic phase predicts where target is more likely to move, and it is based usually on a first- or second-order autoregressive model. 3) The observation phase finds the region of the image that is more similar to the target, assuming as prior the hypothesis given in the dynamical phase.

Tracking, and especially the dynamic module, may benefit from Social Signal Processing methods. Such module simply does not take into account that people, whenever free to move in a large environment (e.g., the hall of a hotel, a square, a waiting room, etc.), respect patterns and trajectories largely dominated by social mechanisms [18]. Therefore, the design of a socially driven dynamic model for tracking may be the key ingredient to overcome the current limitations of the current algorithms, as already shown in some recent approaches exploiting the Social Force Model [19, 20] (see later for further details) . When the scenario is too crowded, so that tracking approaches become ineffective, motion flow estimation techniques are usually preferred [21].

Activity analysis. Activity analysis and recognition[1] is the last module added in the typical scheme of a surveillance system. It usually takes the trajectories of the targets and the object detections results as input, and provides a description of the activities carried out in the scene, under the form of parametric models (in most of the cases, we have a model for each activity) or natural language expressions [22]. The general idea is to segment the trajectories and the detections of each subject into simple actions, by means of actions classifiers that work on few consecutive frames. Then, higher-level reasoning is applied to combine the simple actions through statistical pattern recognition paradigms or methods that exploit temporal logics. It is very important to note that almost all the behavior/activity recognition modules are strongly context dependent: the definition of the plausible simple actions and activities are tightly linked to the kind of monitored environment (an airport, a parking lot). In all the cases, the activity recognition approaches ignore that the human behavior is a process subject to laws rigorous enough to produce stable patterns corresponding to social, emotional, and psychological phenomena. It turns out that this module is the one where social signal processing applies mostly. For this reason we will now give a short overview of what has been done so far for this stage in the "classical" Computer Vision sense, highlighting later the main limitations where social signaling may help the most.

3. Classical activity analysis: a short review

This overview is not exhaustive, and wants only to give an idea on the kind of activities considered so far in surveillance. For more detailed overviews, please consider the surveys referred in the Introduction.

Approaches that follow the scheme of Sec. 2 for individuating single-agent activities, based on tracking trajectories, are [23, 24, 25, 26].

One of the most known classical system for automated surveillance is VSAM [23], where individual activities like entering vehicles, entering buildings etc. are encoded as simple Markov models. Another masterpiece is [24], where expected trajectories are quantized and learnt by neural networks, and the goal is that of

[1] It is worth noting that in the surveillance literature, the definitions of behavior analysis and activity recognition are often used without distinction, assuming that a behavior or activity is an ordered sequence of simple actions (walking, running, stopping, meeting) performed by one or more interacting subjects.

finding abnormal events as outliers. In [25], the clustering is hierarchical, producing a hierarchical binary tree. A finer use of trajectory data is presented in [27, 26], where semantic zones like entry/exit zones, junctions, paths, routes, sink, sources and stop zones are located in the monitored scene.

The use of spatio-temporal features instead of trajectories is gaining more popularity for finely analyzing individual human behavior [28, 29, 30, 31]: such features overcome the limitations of tracking-based schemes by exhibiting robustness to noise, small camera movements and changes in lighting conditions, allowing to encode activities as walking, jumping, bending, turning around, kicking etc. .

Moving to activities involving more than a person, Dynamic Bayesian Networks [32] such as Hidden Markov Models (HMMs) [33] and more complex models which build upon HMMs are the most used tools [34, 35, 36]. In [34], a total of three kinds of activities (following, meet and continue together, meet and go on separately) on a public square are modeled by means of Coupled Hidden Markov Models (CHMM), operating on trajectories. Semi-Markov reasoning for encoding long term activities is proposed in [35]: simple events (running, approaching, etc.) are temporally composed in order to define complex events (a person runs and then slows down). Interactions are modeled by logic operators that assemble together single-thread (i.e., performed by a single person) complex events into a multi-thread complex event. In [36], video sequences are represented at different scales in terms of different motion details related to trajectory, silhouette, body parts, etc. Then, these scales are combined using a hierarchical Directional-State Dynamic Bayesian Network (HDS-DBN) to perform recognition of activities like two people walking in the same or opposite direction, people interaction, dropping and picking up of an object.

A marriage between syntactical and statistical pattern recognition paradigms has been proposed in [37]. A low-level module detects by tracking simple events (for example, enter and move in a parking lot). These events are then fed into a stochastic context-free parser that connects atomic events by exploiting longer range temporal constraints. More recently, the same principle is adopted and developed in [38] to deal with interactive activities as greeting, fighting.

Concerning the approaches based on spatio-temporal features, two-agents interactions are modeled in [39], employing spatial and temporal logic for the classification of activities like shaking hands, hugging, punching, pointing.

All the above approaches focus on activities performed by 1-2 people, under the form of sequences of explicit actions. Interesting aspects as personality traits, or intentions, which could be useful for *predicting* activities, are not taken into account.

The following methods bring into the analysis the concept of group of objects.

In [40], groups are represented by a geometrical shape, in which vertices are the locations assumed by the moving persons along time. The idea is that a "mean shape" represents a particular group activity and variations with respect to it indicate abnormal events. The variations can be spatial (a person is in an unexpected location) or temporal (for example, a person stands still). This system has been applied in an airport scenario, where the interacting people were passengers moving from the airplane to the terminals. Airport cargo loading/unloading activities, structured as multiple interactions between vehicles, with actions like moving truck, moving cargo, are modeled in [41] with Dynamic Bayesian Networks, whose structure was learnt in an automatic fashion. More recently, in [42], group activities are encoded with three types of localized causalities, namely self-causality, pair-causality, and group-causality, which characterize the local interaction/reasoning relations within, between, and among motion trajectories of different humans, respectively. Six different human group activities (8 persons maximum) are considered, i.e., walk-in-group, run-in-group, stand-and-talk, gathering, fighting, and ignoring (i.e., the subjects walk independently). The same authors improved their framework in [43], employing Gaussian Processes for describing motion trajectories. In [44], group interactions with a varying number of subjects are investigated, employing an asynchronous HMM as a hierarchical activity model. They distinguish symmetric (like *i* talks with *j*) and asymmetric dynamics activities (like *i* follows *j*). In particular, they focus on InGroup, Approach, WalkTogether, Split, Ignore, Chase, Fight, and RunTogether activities.

Another generative model is presented in [21], where interacting events in crowded scene are modeled in an unsupervised way, and interactions are modeled as co-occurrences of atomic events. No tracking is performed due to the high density of people, and local motions are considered as low-level features instead. After collecting such motion patterns, atomic events can be defined as distributions over these low-level features and, in the same way, interactions are modeled

as distributions over atomic events. All the distributions are modeled with dual Hierarchical Dirichlet Process, which decides in an automatic fashion both the number of atomic events and interactions. The system works very well in detecting interactions in traffic scenes, with cars and pedestrian, but it seems less expressive in modeling human interactions. The discriminative approach in [45] encodes the context as a mean for inferring individual activities in a more robust way. Two types of contextual information are used: the first captures the main activity performed by a group of people, the second evaluates the close neighborhood of a person. Five actions are considered (crossing, waiting, queuing, walking and talking).

In order to understand the state of the art at a glance, we organize the approaches discussed so far by considering three aspects: the first is the *degree of environmental supervision* (DES), that focuses on how much the monitored scenario is constrained: highly constrained scenarios (DES=H) correspond to small ambients, possibly monitored by several multimodal sensors; viceversa, unconstrained situations occur when large outdoor scenes are captured by a single camera (DES=L). The second aspect is the *level of detail of the interaction* (LDI), which refers to the level of detail with which an interaction is modeled. From one side, we have highly specific interactions (LDI=H), where body gestures are needed; on the other side, we have generic interactions, where each individual is represented as a simple point, whose only position and motion are considered (LDI=L). The third aspect is the *number of subjects* (NOS), that takes into account the number of people involved in an activity. In Table 1, we organize the surveyed papers accordingly to the DES, LDI and NOS characteristics.

From a pure technical point of view, another possible taxonomy can be defined according with the type of methodology employed. To this end, the techniques can be partitioned into three classes, namely: 1) the graphical models-based approaches or, more simply, generative models; 2) the discriminative approaches, 3) the syntactical approaches. The generative models include Markov models [23], Bayesian networks [40], Dynamic Bayesian Networks (mostly Hidden Markov Models) [34, 35, 36, 44, 21, 41, 27, 26], non parametric models [25, 28], generic Bayesian models [30].

The discriminative approaches usually adopt Support Vector Machines [29, 42], with different kinds of kernels [39], Relevance Vector Machines [31], Latent SVMs [45], Gaussian Processes [43], neural networks [24].

The syntactic approaches dictate the construction of grammars which are then used to express the structure of a process using a set of production rules [37, 38].

Table 1: Summary of the reported works on activity recognition

Approach	DES	LDI	NOS
[23, 24, 25, 27, 26]	L	L	=1
[28, 29, 30, 31]	H	H	=1
[34, 37]	L	L	2
[35, 36, 38, 39]	L	H	2
[41, 40, 42, 21, 44, 43]	L	L	>2

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 activity analysis level, much more can be done. In the following, we list different problems where the lack of social knowledge clearly emerges.

- PROBLEM 1:*Definition of threatening behavior* All the surveillance systems aims at promptly identifying threatening behaviors, but it turns that most of them provides "unusual" activities, considering a previously learned statistics. Especially when this statistics collected is scarce, this will cause huge amounts of false positives, making the system unusable for practical purposes. Therefore, a different definition of "threatening" behavior has to be forged.

- PROBLEM 2:*Modeling of groups* What is a group? In the surveillance literature this usually corresponds to having a set of individuals exhibiting similar characteristics, i.e., close in space, with the same oriented motion. This description fails to distinguish a situation where space constraints force people to wander close from standard proximity given by personal relationships. Therefore, a more expressive definition of group has to be designed.

- PROBLEM 3:*Modeling of interactions in outdoor scenarios* 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. In outdoor scenarios, the simple spatial proximity is usually assumed as warrantee for interaction: this is intuitively false in crowded situations like cocktail parties. Therefore, a more precise definition of interaction has to be provided.

As we will see in the following, Social Signal Processing may help in answering the above questions, providing novel cues that can be exploited by standard surveillance algorithms.

4. Social Signal Processing for activity recognition: toward the analysis of the behavior

Social Signal Processing aims at developing theories and algorithms that codify how human beings behave while involved in social interactions, putting together perspectives from sociology, psychology, and computer science [10, 11]. Here, the main tools for the analysis are the social signals [11], *i.e.*, temporal co-occurences of social or behavioral cues [46], that can be basically defined as a set of temporally sequenced changes in neuromuscular, neurocognitive, and neurophysiological activity. Behavioral cues (see Figure 2 for some examples) have been organized into five categories in [11] that are heterogeneous, multimodal aspects of a social interplay: 1) *physical appearance*, 2) *gesture and posture*, 3) *face and eyes behavior*, 4) *vocal behavior* and 5) *space and environment*.

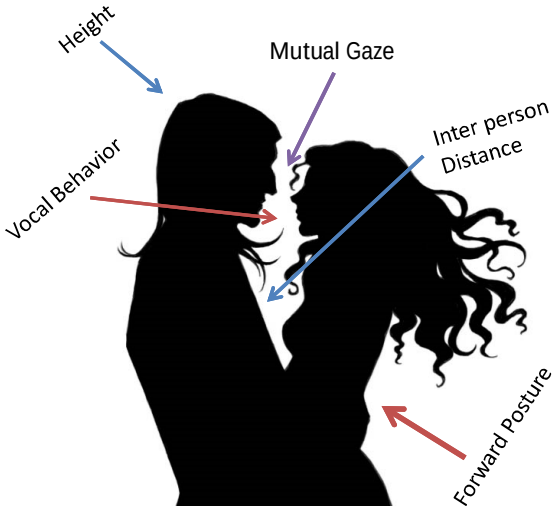

Figure 2: Example of behavioral cues.

The interaction of Social Signal Processing and Automated Surveillance is, to our opinion, at its infancy. Actually, the most SSP approaches deal with well-constrained scenarios as smart or meeting rooms, where the sensor machinery is massive and pervasive. In this way, fine behavioral cues can be captured, especially the ones focusing on the gesture and posture, the face and eyes behavior and the vocal behavior. Such capabilities cannot be exploited in a typical surveillance scenario: actually, smart sensors cannot be placed in the environment because of 1) privacy protection measures, and 2) scarce effectiveness in wide open scenes.

It turns out that few SSP works are concerned with surveillance [11], and they involve primarily the category of cues related to the space and the environment, that in our opinion represents the most intuitive connection. Actually, from the SSP side, this category has been extensively investigated in human sciences, where the spatial arrangement of people in social encounters (also called *proxemics*) has been shown to be a reliable evidence of the social phenomena taking place among interacting individuals [47, 18]. From a surveillance perspective, the encoding of proxemics aspects comes along with the tracking and classification technologies, that provide the relative position of people at each frame.

In the rest of the section, the five categories of behavioral cues are detailed, reviewing the surveillance approaches that explicitly use them, and envisaging possible future perspectives of cross-pollination.

4.1. Physical Appearance

Physical appearance of a person codifies attributes like attraction, height and somatotype. Attractiveness is an important physical factor as it pushes one to interact. Attractive people have a high probability of getting in contact with other people [13]. Research in the area of facial perception has identified many different factors that contribute to a face being considered attractive and it is generally accepted that beauty cannot be defined by one single principle [48]. The second important physical attribute is height: people tend to attribute high social status to taller people. Finally, somatotypes (being tall and thin, proportioned, short and fat) tend to draw some attributes of personality traits. For example, thin people are considered to express less emotion, while fat people tend to be more talkative. To the best of our knowledge, this class of cues has not been exploited in any surveillance approach, and is also absent in the more general literature of the automated analysis of behavior. Therefore, its use for surveillance seems unlikely.

4.2. Gesture and Posture

Gestures are used to regulate the human interaction and they often performed consciously or unconsciously to convey some specific meaning. Furthermore, gestures as in Figure 3(a) can also express an emotion and

hence capable of convey the social signals for the most complicated human behaviors like shame and embarrassment [11]. The usage of the gestures in a social signaling-driven surveillance sense is hard: the goal is not only capturing intentional gestures that are voluntarily expressed by the subjects for communicating something, but it is also capturing unintentional movements, subtle and/or rapid oscillations of the limbs, casual touching of the nose/ear, hair twisting and self protection gesture like closing the arms. These cues are very hard to be modeled since they are inherently affected by noise and occlusions. While the former attempt has been pursued by a huge quantity of works, the goal of capturing *subtle* gestures is a big challenge, with no traces in the literature.

In [49], gesturing is used to infer who is talking when in a surveillance scenario, realizing through statistical analysis a simple form of diarization (detection of *who speaks when*, [50]). Actually, cognitive scientists showed that speech and gestures are so tightly intertwined that every important investigation of language has taken gestures into account [51]. While the multimodal diarization is common in the literature (based on the joint modeling of speech, facial and bodily cues), the unimodal diarization exploiting visual information is rare [52], unrelated to surveillance. We think that this is a direction worth to be investigated, because this allows to capture turn taking patterns indicating ongoing conversations and thus genuine social interactions (PROBLEM 3).

As it can be seen in Figure 3(b), the posture is an aspect of the human behavior which is unconsciously regulated and thus can be considered as the most reliable nonverbal social cue. In general, posture conveys social signals in three different ways [11], namely: inclusive vs. no inclusive, face-to-face vs. parallel body orientation, and congruent vs. non-congruent. These cues may help to distinguish extrovert and introvert individuals, suggesting a mean to individuate threatening behaviors (PROBLEM 1). Only few and very recent surveillance approaches deal with posture information [53, 54, 55, 56]: they will be explained in the following, since they exploit mostly cues coming from other behavioral categories.

4.3. Face and Gaze behavior

These are termed as the best efficient social signals that can describe the human behavior and also have an impact on our perception about others affect [57]. Figure 4 shows the example of facial expressions and gaze direction that can be termed as strong social signals.

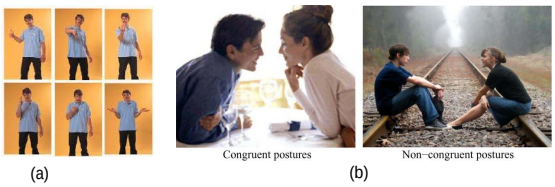

Figure 3: Examples of (a) Gesture (b) Postures (taken from [11])

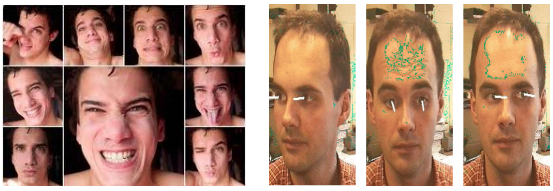

Figure 4: Examples of facial expression and gaze direction

Nonverbal facial cues includes fear, sad, happiness, anger, disgust, surprise, psychological states like suicidal and depression and also social behavior like rapport and accord.

In surveillance, the goal of capturing fine visual cues from the face is very hard, since we are in a non-collaborative scenario (people do not look at the sensors) and the sensors usually capture the faces at a low resolution.

A different matter holds for the gaze orientation. Since objects are foveated for visual acuity, gaze direction generally provides precise information regarding the spatial localization of ones attentional focus [58], also called Visual Focus of Attention (VFOA). Concerning social aspects, VFOA is a fundamental mean of nonverbal communication [59, 60, 61], so that its modeling is very attractive in surveillance. The problem is that measuring the VFOA by using eye gaze is often difficult or impossible in standard surveillance scenarios: either the movement of the subject has to be constrained or high-resolution images of the eyes are required [62]. Therefore, the viewing direction has been reasonably approximated by just measuring the head pose [59, 63].

In such scenario, the work of [61] estimates pan and tilt parameters of the head and the VFOA is represented as a vector normal to the person's face. The application purpose is to understand if a person is looking at an advertisement located on a vertical glass or not. Since the specific setup is very constrained, this model works pretty well. However, as observed by the authors, more general setups impose more complex models that consider camera position, people location and scene structure. Similar considerations hold for the work presented

in [60], where an Active Appearance Model models the face and pose of a person in order to discover which portion of a mall-shelf is observed.

Following this claim, and considering a general, unrestricted scenario, where people can enter, leave, and move freely, the VFOA can be approximated by the three-dimensional (3D) visual field of an individual. More precisely, according to biological evidence [64], the VFOA can be described as a 3D polyhedron delimiting the portion of the scene that the subject is looking at (see Figure 5).

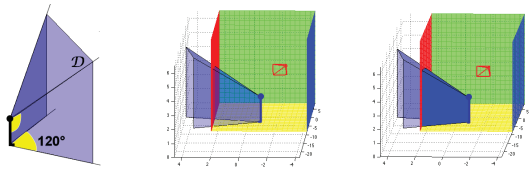

Figure 5: Left: the VFOA model. Center: an example of VFOA inside a 3D "box" scene. In red, the camera position: the VFOA orientation is estimated with respect to the principal axes of the camera. Right: the same VFOA delimited by the scene constraints (in solid blue).

The use of the gaze approximation as a 3D polyhedron in surveillance brought a radical change of perspective for the behavior analysis. We moved from a objective point of view, i.e., the point of view of the surveillance camera, toward a subjective point of view, i.e., that of each single individual. The gaze approximation allows to understand what a person is looking at, building a set of high-level inferences.

For example, in [65], and, independently, in [66] the idea was to infer what part of the scene is seen more frequently by people, thus creating a sort of interest maps. This may serve to highlight individuals that are focused on particular portions of the environment for a long time: if the observed target is critical (for example, an ATM machine) a threatening behavior could be inferred (PROBLEM 1).

Later on, the "subjective" perspective has been proposed in [53], where group interactions are discovered by estimating the visual focus of attention using a head orientation detector, while exploiting proxemic cues. The idea is that close people whose VFOA is intersecting are interacting (PROBLEM 2).

Similarly, in [54], a set of two and one-person activities are formed by sequences of actions and then modeled by HMMs whose parameters are manually set.

The importance of the "subjective" point of view for surveillance encourage scientists in ameliorating the VFOA extraction phase. Most recently, in [67], an approach for the joint tracking in surveillance videos of pose behavioral cues (body position/pose and head pose) is presented. Given the tracks generated by a multi-person tracker, they first localize the head and extract body and head pose features. Then, these features are used to jointly estimate the pose cues in a 3D space using a particle filtering approach that exploits the conditional coupling between body position (movement direction) and body pose together with the soft coupling between body pose and head pose.

Table 2: Summary of SSP approaches for surveillance purposes exploiting the face and gaze behavior cues.

Approach	DES	LDI	NOS
[61, 60]	H	H	1
[65, 66, 67]	L	H	1
[54]	L	H	≤ 2
[53]	L	H	> 2

To summarize, we report the table of all the reviewed approaches exploiting face and gaze behavior cues, considering the three aspects previously described in Sec.3, i.e., DES, LDI, and NOS.

4.4. Vocal behavior

The vocal behavior class comprehends all the spoken cues that define the verbal message and influence its actual meaning. Such class includes five major components [11]: *prosody*, that can provide social signals like competence; *linguistic vocalization*, that can communicate hesitation; *non linguistic vocalization*, that can provide strong emotional states or tight social bonds, *silence*, that can express hesitation, and *turn taking patterns*: this last component is the most investigated in this category, since it appears the most reliable when the goal is to recognize people personality [68], predict the outcome of negotiations [69], recognize the roles interaction participants play [70], or modeling the type of interactions (e.g., a conflict).

As turn-organization cannot be fully understood without taking into account its sequential aspects [71], the application of probabilistic sequential models is widespread. In [72], a two-layer HMM was employed to model individual and group actions (e.g., discussions, presentations, etc.). In [73], the purpose was to detect the dominant interlocutor through social cues of mimicking. The authors employed an *Observed Influence Model* (OIM), i.e., an aggregate of first-order Markov processes, each one addressing an interlocutor.

More recently [74, 75], a generative framework has been proposed aimed at classifying conversation intervals of variable length (from a few minutes to hours),

considering the nature of the people involved within (children, adults) and the main mood.

In surveillance, approaches that face monitoring scenarios considering vocal behavior cues are absent, since the audio modality is hard to be captured in wide areas, and, most importantly, it is usually forbidden for privacy issues. The problem lies in the fact that audio processing is usually associated with speech recognition, while in the case of SSP, the content of a conversation is ignored.

An interesting topic for surveillance is that of the modeling of conflicts, as they may degenerate in threatening events (PROBLEM 1). Conflicts have been studied extensively in a wide spectrum of disciplines, including Sociology (e.g., see [76] for social conflict) and Social Psychology (e.g., see [77] for intergroup conflicts).

An approach that could be instantiated in a surveillance context is that of [78], that proposes a semi-automatic generative model for the detection of conflicts in conversations. The approach is based on the fact that, during conflictual conversations, overlapping speech becomes both longer and more frequent [79], the consequence of a competition for holding the floor and preventing others from speaking.

In summary, vocal behavior appears to be a very expressive category of social cues, that should be exploited in the surveillance realm, since it can be handled in a completely privacy-respectful fashion.

4.5. Space and Environment

The study of the space and environment cues is tightly connected with the concept of proxemics that can be defined as the "[...] *the study of man's transactions as he perceives and uses intimate, personal, social and public space in various settings* [...]", quoting Hall [18], 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. This happens 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 [80]), attitudes (e.g., people that discuss tend to seat in front of the other, whereas people that collaborate tend to seat side-by-side [81]), etc. . From a social point of view, two aspects of proxemic behavior appear to be particularly important, namely interpersonal distances and spatial arrangement of interactants.

Interpersonal distances have been the subject of the earliest investigations on proxemics and one of the main and seminal findings is that people tend to organize the space around them in terms of four concentric zones associated to different degrees of intimacy:

Intimate Zone: distances for unmistakable involvement with another body (lover or close friend). This zone is typically forbidden to other non-intimate persons, except in those situations where intrusion cannot be avoided (e.g. in elevators).

Casual-Personal Zone: distances established when interacting with familiar people, such as colleagues or friends. This zone is suitable for having personal conversations without feeling hassled. It also reflects mutual sympathy.

Socio-Consultive Zone: distances for formal and impersonal relationships. In this zone, body contact is not possible anymore. It is typical for business conversations, consultation with professionals (lawyers, doctors, officers, etc.) or seller-customer interactions.

Public zone: distances for non-personal interaction with others. It is a zone typical for teachers, speakers in front of a large audience, theater actors or interpersonal interactions in presence of some physical barrier.

In the case of Northern Americans, the four zones above correspond to the following ranges: less than 45 *cm* (intimate), between 45 and 120 *cm* (casual-personal), between 120 and 200 *cm* (socio-consultive), and beyond 200 *cm* (public). While the actual distances characterizing the zones depend on a large number of factors, as we will see in the following, the partition of the space into concentric areas seems to be common to all situations.

The spatial arrangement during social interactions addresses two main needs: the first is to give all people involved the possibility of participating, the second is to separate the group of interactants from other individuals (if any). The result are the *F-formations*, stable patterns that people tend to form during social interactions (including in particular standing conversations): "*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*" [82].

In practice, an F-formation is the proper organization of three social spaces (see Figure 6 (a)): O-space, P-space and R-space. The O-space (the most important component of an F-formation) is a convex empty space surrounded by the people involved in a social interaction, every participant looks inward into it, and no external people are allowed in this region. The P-space is a narrow stripe that surrounds the O-space and that contains the bodies of the interactants, the R-space is the

area beyond the P-space.

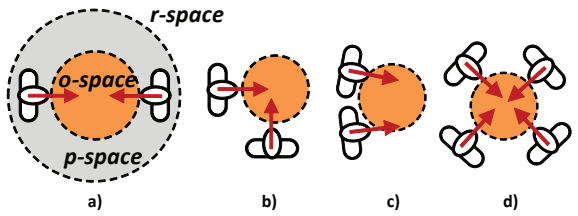

Figure 6: F-formations; a) scheme of a (vis-a-vis) F-formation; b) L-shape; c) side-by-side; d) circle.

There can be different F-formations (Figure 6 (b)-(d)):

Vis-à-vis: An F-formation in which the absolute value of the angle between participants is approximately 180^o, and both participants share an O-space.

L-shape: An F-formation in which the absolute value of the angle between participants is approximately 90^o, and both participants share an O-space.

Side-by-side: An F-formation in which the absolute value of the angle between participants is approximately 0^o, and both participants share an O-space.

Circle: An F-formation where people is organized in a circle, so that the configuration between adjacent participants can be considered as a hybrid between a L-shape and a Side-by-side F-formation.

The proxemic behavior intended as the use of the interpersonal distances is affected by a large number of factors, and culture seems to be one of the most important ones, especially when it comes to the size of the four concentric zones described above. In particular, cultures seem to distribute along a continuum ranging from "contact" (when the size of the areas is smaller) to "non-contact" (when the size of the areas is larger) [18]. The effect of culture seems to change when interaction participants have seats at disposition. In this case, people from supposedly "non-contact" cultures tend to seat closer than the others [83].

Interesting effects have been observed considering the size and the illumination of the interaction site: people allow others to come closer in larger rooms [84], in bright ambients [85], when the ceiling is higher [86], and in outdoor spaces [87]. The effects of crowding have been studied as well [88]: social density was increased in a constant size environment for a limited period of time and participants of larger groups reported greater degrees of discomfort and manifested other forms of stress.

To the best of our knowledge, only a few works have tried to apply proxemics in computing. One probable reason is that current works on analysis of human behavior have focused on scenarios where proxemics do not play a major role or have relied on laboratory settings that impose too many constraints for spontaneous proxemic behavior to emerge (e.g., small groups in smart meeting rooms) [89]. Most of the computing works that can be said to deal with proxemics concern the dynamics of people moving through public spaces.

The keystone model for the interaction modeling of moving people, i.e. a basic form of behavior, is given in the social force model (SFM) [90], that applies a gas-kinetic analogy to the dynamics of pedestrians. It is not a interaction detection system, rather it is a physical model for simulating pedestrian interactions while they are moving. Pedestrians are assumed to react to energy potentials caused by other pedestrians and static obstacles through a repulsive or an attractive force, while trying to keep a desired speed and motion direction. This model can be thought as explaining group formations, and obstacle avoidance strategies, i.e., basic and generic form of human interactions. The social force model has been modified in [19], where SFM is embedded in a tracking framework, substituting the actual position of the pedestrian of the SFM with a prediction of the location made by a constant velocity model, which is then revised considering repulsive effects due to pedestrians or static obstacles. No mention about attractive factors are cited in the paper. At the same time, independently, a variational learning strategy is proposed in [20] to train a dynamic model for predicting the position of moving subjects, employing the SFM. Even in this case, the attraction factor of the SFM is ignored.

In [91], a versatile synergistic framework for the analysis of multi-person interactions and activities in heterogeneous situations is presented. An adaptive context switching mechanism is designed to mediate between two stages, one where the body of an individual can be segmented into parts, and the other facing the case where persons are assumed as simple points. The concept of spatio-temporal personal space is also introduced to explain the grouping behavior of people. They extend the notion of *personal space* to that of *spatio-temporal personal space*. Personal space is the region surrounding each person, that is considered personal domain or territory. Spatio-temporal personal space takes into account the motion of each person, modifying the geometry of the personal space into a sort of cone. Such cone is narrowed down proportionally with the motion

of the subject so as the faster the subject, the narrower the area. An interaction is then defined as caused by intersections of such volumes.

In [55], F-formations are found in a cocktail party scenario by employing proxemics elements and head orientation estimates. The approach is based on a Hough voting strategy, and represents an accurate modeling of the formal definition of F-formation. The main characteristics are that people have to be reasonably close to each other, have to be oriented toward the o-space, and that the O-space has to be empty to allow the individuals to look at each other.

Another approach for the F-formation is that of [56]. They define an F-formation as a set of focused encounters, and this distinction serves to discriminate a group where people is willing to stay to those group formations resulting from environmental constraints (people that move in a narrow road). The authors use a graph clustering algorithm by formulating the problem in terms of identifying dominant sets. A dominant set is a form of maximal clique which occurs in edge weighted graphs. As well as using the proximity between people, body orientation information is used.

These two last approaches seem to be particularly indicated to highlight genuine group formations and interactions, where proxemic cues and postural cues go beyond the mere spatial proximity exploited in most of the cases (PROBLEM 2,3).

In [92], the authors propose a system that could track and discover groups of interacting people, estimating the trajectories of people and employing the modularity cut algorithm [93]. A limitation of the work was that of considering solely staged social activities.

One of the first attempts to interpret the movement of people in social terms has been presented in [94], where nine subjects were left free to move in a $3m \times 3m$ area for 30 minutes. The subjects had to speak among themselves about specific themes. An analysis of mutual distances in terms of the zones described in Section 4.5 allowed to discriminate between people who did interact and people who did not.

In a similar way, mutual distances have been used to infer personality traits of people left free to move in a room [95]. The results show that it is possible to predict Extraversion and Neuroticism ratings based on velocity and number of intimate/personal/social contacts (in the sense of Hall) between pairs of individuals looking at one other.

The approach of [96] studies social relations in F-formation, calculating pairwise distances between people lying in the p-space, and clustering them in different classes. The number of classes is chosen automatically by the algorithm, following a Information Theory principle. The main finding of the approach is that each of the classes actually represent well-defined social bonds. In addiction, the approach adapts to different environmental conditions, namely, the size of the space where people can move.

Table 3: Summary of SSP approaches for surveillance purposes exploiting the space and environment cues.

Approach	DES	LDI	NOS
[90, 19, 20]	L	L	> 2
[91]	L	L/H	> 2
[55, 56, 92, 96]	L	H	> 2
[94, 95]	H	H	> 2

The last three approaches are very close to a modelling of social roles in unconstrained scenarios, that in turn may serve to the detection of threatening behavior (PROBLEM 1).

Finally, we report a concise scheme of the reviewed approaches in Table 3 exploiting space and environment cues in a social signaling sense for surveillance purposes, still considering the three aspects of DES, LDI, and NOS.

5. Crowd behavior analysis

Analyzing a crowd represents without doubts a new dimension for the automated surveillance. The idea is to monitor huge masses of people, categorizing how they move, and looking for normal and abnormal situations, due for example to incidents, panic attacks etc. . This mission is intriguing, because it requires to revise the whole surveillance flowchart previously described in Section 2. Actually, each single individual here cannot be characterized as finely as in the case of 5–10 people: occlusions are very strong, and the classical object classification and tracking approaches have shown to be scarcely effective [97]. The idea is that there are no more many entities to model, but instead a single one, the crowd. The underlying hypotheses are that a crowd has an own appearance to be modeled, it moves with a very complex dynamics that can be learned, and behaves following sociological principles.

Crowd analysis was born in the field of transportation and public safety [97, 98], considering three main applications: (a) density estimation (counting individuals from crowd); (b) tracking some individuals in crowded scene; (c) crowd behavior understanding. The most studied model in crowd behavior analysis is the social force model (SFM) already introduced in the previous Section 4 about proxemics.

Many variants have been proposed, still for simulation targets, but, recently, some works suggest to invert this model to detect and localize the abnormalities present in the crowd [99]. The use of SFM in fact avoids individual person tracking and this is the key aspect of its success. Other methods have been subsequently proposed having this same characteristic.

In [99], grid of particles is placed over the image and these particles are advected with the space-time average of optical flow. Then, the interaction force for each particle is estimated using SFM. These interaction forces are then mapped onto the image plane to obtain a force flow for every pixel in every frame. The resulting vector field is used to model the normal behavior using a bag of words approach.

In [98], an individual target tracking strategy for unstructured crowded scenes is presented. Here, the crowd is modeled using Correlation Topic Model (CTM): the idea is that the dynamics of the crowd is learned by quantized local features, correlated together with the use of topic models. Given the observed measurements of the object to be tracked, the next position is obtained by incorporating in the object state hypothesis the learned high-level scene dynamics.

In [100], a tracking scheme based on local spatio-temporal motion patterns is presented to track an individual in an extremely crowded scene. Spatio-temporal variations of the crowd motion are learned from regularly spaced small subvolumes of the considered video sequence using a battery of HMMs. These HMMs are subsequently used to predict the motion patterns of a given subject when deviating from the main crowd flow.

A new scheme to detect the motion patterns in the crowded scenarios is presented in [101]. This scheme utilizes the instantaneous motion flow estimating the optical flow field. The typical motion patterns are then detected by clustering the flow vectors from the motion flow.

Recently, another approach has been proposed still inspired by the SFM and its variants [99]. These works utilize a set of particles over each frame and the SFM formulation to estimate the force of each particle. Subsequently, the set of particles' forces has been optimised using a swarm optimization technique so that abnormal situations can be detected in a faster and more reliable way, also localizing the anomalies [102, 103].

At the best of our knowledge, current methods are still far from coupling surveillance and social signal processing in the field of crowd analysis. Under a genuine sociologic point of view, an important survey that deal with the modelling of crowd, particularly suited for the field of public transportation and safety is [104]. Here, different theories like Pre-disposition theory, Emergent Norm theory, Model of Disorder, Social Identity theory and Elaborate social identity theory are discussed. This could be a point where Computer Vision may draw on. The idea is that the coupling of sociological notions and computer vision algorithms may originate novel applications. For example in the following, we foresee the development of two possible activities:

- *Design of public spaces*. Simple architectural elements are known to influence significantly the collective behavior of large crowds in public spaces [105]. 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 a large number of interacting individuals.

- *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 [106]. 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.

6. Conclusions and future research

The technical quality of the classical modules that compose a surveillance system allows nowadays to face very complex scenarios. The goal of this review is to support the argument that a social perspective is fundamental to deal with the highest level module, i.e. the analysis of human activities, in a principled and fruitful way. We discussed how the use of social signals may be valuable toward a robust encoding of social events that otherwise cannot be captured. In particular, we indicate three problems, that is, the definition of threatening behavior, the modeling of groups and the modeling of interactions in outdoor situations, that are very frequent in video surveillance and that can be faced with more effectiveness under a Social Signal Processing perspective. In short, Computer Vision and Pattern Recognition furnish the analysis tools to be exploited following social science findings. We are convinced that this is the way surveillance expressiveness may be boosted, leaning toward a finer investigation of overt and covert, also subtle, human behavioral aspects. In addition, importing social models into crowd analysis represents a very

fertile and still unexplored area where many contributions could be provided.

References

[1] C. Cedras, M. Shah, Motion-based recognition: A survey, Image and Vision Computing 13 (2) (1995) 129 – 155.

[2] J. K. Aggarwal, Q. Cai, Human motion analysis: A review, Computer Vision and Image understanding 73 (3) (1999) 428 – 440.

[3] D. M. Gavrila, The visual analysis of human movement: A survey, Computer Vision and Image understanding 73 (1) (1999) 82 – 98.

[4] T. B. Moeslund, A. Hilton, V. Krger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image understanding 104 (2) (2006) 90 – 126.

[5] H. Buxton, Learning and understanding dynamic scene activity: a review, Image Vision Computing 21 (1) (2003) 125–136.

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

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

[8] H. Dee, S. Velastin, How close are we to solving the problem of automoted visual surveillance, Machine vision and application 19 (2) (2008) 329 – 343.

[9] J. Aggarwal, M. Ryoo, Human activity analysis: A review, ACM Computing Survey 43 (2011) 1–43.

[10] A. Pentland, Social signal processing, IEEE Signal Processing Magazine 24 (4) (2007) 108 – 111.

[11] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: Survey of an emerging domain, Image and Vison Computing 27 (2) (2009) 1743 – 1759.

[12] M. Cristani, V.Murino, A. Vinciarelli, Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space, in: Proc. of Workshop on Socially Intelligent Surveillance and Monitoring SISM 2010, 2010, pp. 51 – 58.

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

[14] M. Cristani, M. Farenzena, D. Bloisi, V. Murino, Background subtraction for automated multisensor surveillance: A comprehensive review, EURASIP Journal on Advances in Signal Processing 2010 (2010) 1 – 24.

[15] C. . Conaire, N. E. O'Connor, E. Cooke, A. E. Smeaton, Multispectral object segmentation and retrieval in surveillance video, in: Proc. of International Conference on Image Processing (ICIP), 2006, pp. 8 – 11.

[16] L. Zhang, S. Li, X. Yuan, S. Xiang, Real-time object classification in video surveillance based on appearance learning, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition,(CVPR 2007), 2007, pp. 1 –8.

[17] A. Yilmaz, O. Javed, M. Shah, Object tracking: A survey, ACM Computing Surveys 38 (4) (2006) 1 – 45.

[18] R. Hall, The hidden dimension, 1966.

[19] S. Pellegrini, A. Ess, K. Schindler, L. V. Gool, You'll never walk alone: modeling social behavior for multi-target tracking, in: Proc. of International Conference on Computer Vision, 2009, pp. 261 –268.

[20] P. Scovanner, M. Tappen, Learning pedestrian dynamics from the real world, in: Proc. of International Conference on Computer Vision, 2009, pp. 381–388.

[21] X. Wang, X. Ma, W. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (3) (2009) 539 555.

[22] S. Tellex, D. Roy, Towards surveillance video search by natural language query, in: Proc. of International Conference on Image and Video Retrieval, 2009, pp. 1–8.

[23] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, A system for video surveillance and monitoring, Tech. Rep. CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University (2000).

[24] N. Johnson, D. Hogg, Learning the distribution of object trajectories for event recognition, Image and Vision Computing 14 (1996) 609–615.

[25] C. Stauffer, W. Grimson, Learning patterns of activity using real-time tracking, IEEE Trans. Pattern Analysis Machine Intelligence 22 (8) (2000) 747–757.

[26] D. Makris, T. Ellis, Learning semantic scene models from observing activity in visual surveillance, IEEE Transactions on Systems, Man, and Cybernetics, Part B 35 (3) (2005) 397–408.

[27] C. Stauffer, Estimating tracking sources and sinks, in: Proc. of Computer Vision and Pattern Recognition Workshop (CVPRW), 2003, pp. 35–35.

[28] I. Laptev, T. Lindeberg, Space-time interest points, in: Proc. of IEEE International Conference on Computer Vision, 2003, pp. 432–439.

[29] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proc. of International Conference on Computer Communications and Networks, 2005, pp. 65–72.

[30] P. Yan, S. Khan, M. Shah, Learning 4d action feature models for arbitrary view action recognition, in: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition, 2008.

[31] A. Oikonomopoulos, I. Patras, M. Pantic, Spatiotemporal localization and categorization of human actions in unsegmented image sequences, IEEE Transactions on Image Processing 20 (4) (2011) 1126–1140.

[32] M. I. Jordan, Learning in graphical models, MIT Press, 1999.

[33] L. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, in Proc. of IEEE 77 (2) (1989) 257–286.

[34] N. Oliver, B. Rosario, A. Pentland, Graphical models for recognizing human interactions, in: Proc. of Advances in Neural Information Processing Systems, 1998, pp. 924–930.

[35] S. Hongeng, R. Nevatia, F. Brémond, Video-based event recognition: activity representation and probabilistic recognition methods, Computer Vision and Image Understanding 96 (2) (2004) 129–162.

[36] F. Chen, W. Wang, Activity recognition through multi-scale dynamic bayesian network, in: Proc. of International Conference on Virtual Systems and Multimedia (VSMM), 2010, p. 34 41.

[37] Y. A. Ivanov, A. F. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE Transactions on Pattern Analysis Machine Intelligence 22 (2000) 852–872.

[38] M. Ryoo, J. Aggarwal, Semantic representation and recognition of continued andrecursive human activities, International Journal of Computer Vision 82 (2009) 1–24.

[39] M. S. Ryoo, J. K. Aggarwal, Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in: Proc. of International Conference on Computer Vision (ICCV), 2009, p. 1593 1600.

[40] N. Vaswani, A. R. Chowdhury, R. Chellappa, Activity recognition using the dynamics of the configuration of interacting

objects, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2003, pp. 633–640.

[41] S. Gong, T. Xiang, Recognition of group activities using dynamic probabilistic networks, in: Proc. of nternational Conference on Computer Vision, 2003, pp. 742 – 742.

[42] B. Ni, S. Yan, A. Kassim, Recognizing human group activities with localized causalities, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1470–1477.

[43] Z. Cheng, L. Qin, Q. Huang, S. Jiang, Q. Tian, Group activity recognition by gaussian processes estimation, in: Proc. of International Conference on Pattern Recognition (ICPR), 2010, pp. 3228 –3231.

[44] W. Lin, M. Sun, R. Poovendran, Z. Zhang, Group event detection with a varying number of group members for video surveillance, IEEE Transactions on Circuits and Systems for Video Technology 20 (8) (2010) 1057 –1067.

[45] T. Lan, Y. Wang, W. Yang, G. Mori, Beyond actions: Discriminative models for contextual group activities, in: proc. of Advances in Neural Information Processing Systems (NIPS), 2010.

[46] N. Ambady, R. Rosenthal, Thin slices of expressive behavior as predictors of interpersonal consequences: A meta analysis, Psychological bulletin 111 (2) (1992) 256–274.

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

[48] J. Armstrong, The Secret Power of Beauty: Why Happiness is in the Eye of the Beholder, Penguin Global, London (UK), 2005.

[49] M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, V. Murino, Look at who's talking: Voice activity detection by automated gesture analysis, in: Proc. of Workshop on Interactive Human Behavior Analysis in Open or Public Spaces (InterHub 2011), 2011.

[50] H. Hung, Y. Huang, C. Yeo, D. Gatica-Perez, Associating audio-visual activity cues in a dominance estimation framework, in: Proc. of IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2008.

[51] A. Kendon, Gesticulation and speech: Two aspects of the process of utterance, The Relationship of verbal and nonverbal communication (1980) 207–227.

[52] H. Hung, S. O. Ba, Speech/non-speech detection in meetings from automatically extracted low resolution visual features, in: Proc. of International Conference on Acoustics, Speech, and Signal Processing, 2010, pp. 830–833.

[53] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, V. Murino, Prai*hba special issue: Social interactions by visual focus of attention in a three-dimensional environment, Expert Systems, The Journal of Knowledge EngineeringIn print.

[54] N. Robertson, I. Reid, Automatic reasoning about causal events in surveillance video, EURASIP Journal on Image and Video Processing 2011 1 – 19.

[55] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, A. D. Bue, D. Tosato, G. Menegaz, V. Murino, Social interaction discovery by statistical analysis of f-formations, in: Proc. of British Machine Vision Conference, 2011.

[56] H. Hung, B. Krose, Detecting f-formations as dominant sets, in: Proc. of International Conference on Multimodal Interaction, 2011.

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

[58] S. Ba, J. Odobez, A study on visual focus of attention recognition from head pose in a meeting room, in: Proc. of Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 2006, pp. 75–87.

[59] R. Stiefelhagen, M. Finke, J. Yang, A. Waibel, From gaze to focus of attention, in: Proc. of International Conference on Visual Information and Information Systems, 1999, pp. 761–768.

[60] X. Liu, N. Krahnstoever, Y. Ting, P. Tu, What are customers looking at?, in: Proc. of Advanced Video and Signal Based Surveillance, 2007, pp. 405–410.

[61] K. Smith, S. Ba, J. Odobez, D. Gatica-Perez, Tracking the visual focus of attention for a varying number of wandering people, IEEE PAMI 30 (7) (2008) 1–18.

[62] Y. Matsumoto, T. Ogasawara, A. Zelinsky, Behavior recognition based on head-pose and gaze direction measurement, in: Proc. of International Conference on Intelligent Robots and Systems, 2002.

[63] R. Stiefelhagen, J. Yang, A. Waibel, Modeling focus of attention for meeting indexing based on multiple cues, IEEE Transactions on Neural Networks 13 (2002) 928–938.

[64] J. Panero, M. Zelnik, Human Dimension and Interior Space : A Source Book of Design, 1979.

[65] M. Farenzena, A. Tavano, L. Bazzani, D. Tosato, G.Paggetti, G. Menegaz, V. Murino, M.Cristani, Social interactions by visual focus of attention in a three-dimensional environment, in: Proc. of Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis (PRAI*HBA), 2009, pp. 1 – 8.

[66] B. Benfold, I. Reid, Guiding visual surveillance by tracking human attention, in: Proc.of British Machine Vision Conference, 2009.

[67] C. Chen, A. Heili, J.-M. Odobez, A joint estimation of head and body orientation cues in surveillance video, in: Proc. of IEEE International Workshop on Socially Intelligent Surveillance and Monitoring, 2011.

[68] F. Pianesi, N. Mana, A. Ceppelletti, B. Lepri, M. Zancanaro, Multimodal recognition of personality traits in social interactions, in: Proc. of International Conference on Multimodal Interfaces, 2008, pp. 53–60.

[69] J. Curhan, A. Pentland, Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first five minutes, Journal of Applied Psychology 92 (3) (2007) 802–811.

[70] H. Salamin, S. Favre, A. Vinciarelli, Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction, IEEE Transactions on Multimedia, to appear 11 (7) (2009) 1373–1380.

[71] J. Bilmes, The concept of preference in conversation analysis, Language in Society 17 (2) (1988) 161–181.

[72] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, G. Lathoud, Modeling individual and group actions in meetings with layered HMMs, IEEE Transactions on Multimedia 8 (3) (2006) 509–520.

[73] S. Basu, T. Choudhury, B. Clarkson, A. Pentland, Learning human interaction with the influence model, Tech. Rep. 539, MIT MediaLab (2001).

[74] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, V. Murino, Generative modeling and classification of dialogs by a low-level turn-taking feature, Pattern Recognition 44 (8) (2011) 1785–1800.

[75] A. Pesarin, P. Calanca, V. Murino, M. Cristani, A generative score space for statistical dialog characterization in social signalling, in: Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science 6218, 2010, pp. 630–639.

[76] A. Oberschall, Theories of social conflict, Annual Review of

Sociology 4 (1978) 291–315.
[77] H. Tajfel, Social psychology of intergroup relations, Annual Review of Psychology 33 (1982) 1–39.
[78] A. Pesarin, M. Cristani, V. Murino, A. Vinciarelli, Conversation analysis at work: Detection of conflict in competitive discussions through semi-automatic turn-organization analysis, Cognitive Processing.
[79] E. Schegloff, Overlapping talk and the organisation of turn-taking for conversation, Language in Society 29 (1) (2000) 1–63. doi:10.1017/S0047404500001019.
[80] D. Lott, R. Sommer, Seating arrangements and status., Journal of Personality and Social Psychology 7 (1) (1967) 90–95.
[81] N. Russo, Connotation of seating arrangements, The Cornell Journal of Social Relations 2 (1) (1967) 37–44.
[82] A.Kendon, Conducting Interaction: Patterns of behavior in focused encounters, Cambridge University Press, 1990.
[83] S. Heshka, Y. Nelson, Interpersonal speaking distance as a function of age, sex, and relationship, Sociometry 35 (4) (1972) 491–498.
[84] M. J. White, Interpersonal distance as affected by room size, status, and sex, The Journal of Social Psychology 95 (2) (1975) 241 – 249.
[85] L. Adams, D. Zuckerman, The effects of lighting conditions on personal space requirement, Journal of general psychology 118 (4) (1991) 335–340.
[86] D. Cochran, C, S. Urbanczyk, The effect of availability of vertical space on personal space, Journal of psychology 111 (1982) 137–140.
[87] D. Cochran, C, Personal space requirements in indoor versus outdoor locations, Journal of psychology 117 (1984) 121–123.
[88] W. Griffitt, R. Veitch, Hot and crowded: Influences of population density and temperature on interpersonal affective behavior, Journal of Personality and Social Psychology 17 (1971) 92–98.
[89] D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image and Vision Computing 27 (12) (2009) 1775–1787.
[90] D. Helbing, P. Molnár, Social force model for pedestrian dynamics, Physical Review E 51 (5) (1995) 4282–4287.
[91] S. Park, M. Trivedi, Multi-person interaction and activity analysis: a synergistic track- and body-level analysis framework, Machine Vision Applications 18 (2007) 151–166.
[92] T. Yu, S.-N. Lim, K. Patwardhan, N. Krahnstoever, Monitoring, recognizing and discovering social networks, Proc. of IEEE Conference on Computer Vision and Pattern Recognition.
[93] M. E. J. Newman, Modularity and community structure in networks, Proceedings of the National Academy of Sciences 103 (23) (2006) 8577–8582.
[94] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, L. Schwarz, Detecting social situations from interaction geometry, in: Proc. of IEEE International Conference on Social Computing.
[95] G. Zen, B. Lepri, E. Ricci, O. Lanz, Space speaks: towards socially and personality aware visual surveillance, in: Proc. of ACM international workshop on Multimodal pervasive video analysis.
[96] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, V. Murino, Towards computational proxemics: Inferring social relations from interpersonal distances, in: Proc. of IEEE International Conference on Social Computing, 2011.
[97] J. C. S. J. Junior, S. R. Musse, C. R. Jung, Crowd analysis using computer vision techniques: A survey, IEEE Signal Processing Magazine 27 (5) (2010) 66 77.
[98] M. Rodriguez, S. Ali, T. Kanade, Tracking in unstructured crowded scenes, in: Proc. of IEEE Conference on Computer Vision, 2009, p. 1389 1396.
[99] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935 – 942.
[100] L. Kratz, K. Nishino, Tracking with local spatio-temporal motion patterns in extremely crowded scenes, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 693 – 700.
[101] M. Hu, S. Ali, M. Shah, Learning motion patterns in crowded scenes using motion flow field, in: Proc. of International Conference on Pattern Recognition (ICPR), 2008, pp. 1051–4651.
[102] R. Raghavendra, A. D. Bue, M. Cristani, V. Murino, Optimizing interaction force for global anomaly detection in crowded scenes, in: Proc. of IEEE Workshop on Modeling, Simulation and Visual Analysis of Large Crowds (MSVLC-2011), 2011.
[103] R. Raghavendra, A. D. Bue, M. Cristani, V. Murino, Abnormal crowd behavior detection by social force optimizatios, in: Proc. of Human Behavior Understanding (HBU-2011)), 2011.
[104] K. M.Zeitz, H. M. Tan, M. Grief, P. Couns, C. J.Zeitz, Crowd behavior at mass gatherings: A literature review, Prehospital and Disaster Medicine 24 (1) (2010) 32 38.
[105] P. Ball, Critical mass: How one thing leads to another, William Heinemann Ltd, 2004.
[106] N. Chism, D. Bickford, The importance of physical space in creating supportive learning environments, Jossey-Bass Inc Pub, 2002.