IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 10, 2020, accepted July 5, 2020, date of publication July 10, 2020, date of current version July 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3008370

The Visual Social Distancing Problem

MARCO CRISTANI[1,2,*], (Member, IEEE), ALESSIO DEL BUE[3,*], (Member, IEEE), VITTORIO MURINO[1,3,4,*], (Senior Member, IEEE), FRANCESCO SETTI[1,*], (Member, IEEE), AND ALESSANDRO VINCIARELLI[5,*], (Member, IEEE)

[1]Department of Computer Science, University of Verona, 37134 Verona, Italy
[2]Humatics S.r.l, 37134 Verona, Italy
[3]Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), 16152 Genova, Italy
[4]Ireland Research Centre, Huawei Technologies Company Ltd., Dublin 2, D02 R156, Ireland
[5]School of Computing Science, University of Glasgow, Glasgow G12 8QQ, U.K.

Corresponding author: Alessio Del Bue (alessio.delbue@iit.it)

Alessandro Vinciarelli, Alessio Del Bue, Francesco Setti, Marco Cristani, and Vittorio Murino contributed equally to this work.

This work was supported in part by the projects of the Italian Ministry of Education, United Kingdom Research and Innovation (MIUR), Dipartimenti di Eccellenza, from 2018 to 2022. The work of Alessandro Vinciarelli was supported in part by the United Kingdom Research and Innovation (UKRI) under Grant EP/S02266X/1, and in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/N035305/1.

ABSTRACT One of the main and most effective measures to contain the recent viral outbreak is the maintenance of the so-called Social Distancing (SD). To comply with this constraint, governments are adopting restrictions over the minimum inter-personal distance between people. Given this actual scenario, it is crucial to massively measure the compliance to such physical constraint in our life, in order to figure out the reasons of the possible breaks of such distance limitations, and understand if this implies a potential threat. To this end, we introduce the Visual Social Distancing (VSD) problem, defined as the automatic estimation of the inter-personal distance from an image, and the characterization of related people aggregations. VSD is pivotal for a non-invasive analysis to whether people comply with the SD restriction, and to provide statistics about the level of safety of specific areas whenever this constraint is violated. We first point out that measuring VSD is not only a geometrical problem, but it also implies a deeper understanding of the social behaviour in the scene. The aim is to truly detect potentially dangerous situations while avoiding false alarms (e.g., a family with children or relatives, an elder with their caregivers), all of this by complying with current privacy policies. We then discuss how VSD relates with previous literature in Social Signal Processing and indicate a path to research new Computer Vision methods that can possibly provide a solution to such problem. We conclude with future challenges related to the effectiveness of VSD systems, ethical implications and future application scenarios.

INDEX TERMS Social signal processing, proxemics, human behaviour, person detection, group detection, single view metrology.

I. INTRODUCTION

Humans are social species as demonstrated by the fact that in everyday life people continuously interact with each other to achieve goals, or simply to exchange states of mind. One of the peculiar aspects of our social behavior involves the geometrical disposition of the people during an interplay, and in particular regards the interpersonal distance, which is also heavily dependent on cultural differences. However, the recent pandemic emergency has affected exactly these aspects, as the extraordinary capability of COVID-19 coronavirus of transferring between humans has imposed a sharp and sudden change to the way we approach each other, as well as rigid constraints on our inter-personal distance.

This recently imposed restriction is widely, but imprecisely, referred to as "*social distancing*" (SD) since prevention of the virus diffusion does not require us to weaken our social bonds. The likely reason of SD naming is that, from a cognitive point of view, physical and social aspects of distance are deeply intertwined [47], a phenomenon that popular wisdom captures through a proverb that, in slightly different versions, appears in different languages and cultures, namely "*far from eyes, far from heart*".

Not surprisingly, the time spent in physical proximity with others, in opposition to the time spent in individual activities, is a crucial factor in the "*social brain hypothesis*", one of the most successful theories of human evolution [26].

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

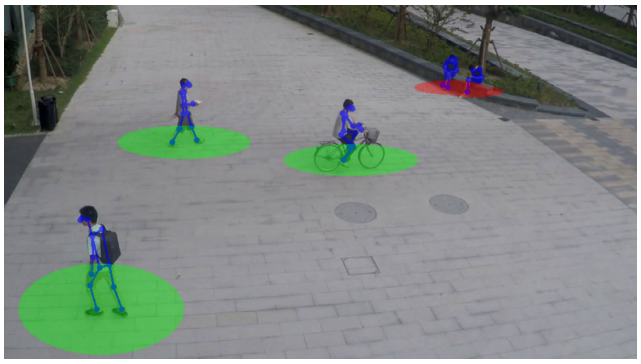


FIGURE 1. The VSD can be estimated in a single frame as the interpersonal distance between people (a 1m radius disk in this example). People that are closer than the imposed distance (red disks), *i.e.* not respecting geometry, might still respect public rules if having a bond of kinship. Results obtained using the following code: https://github.com/IIT-PAVIS/Social-Distancing.

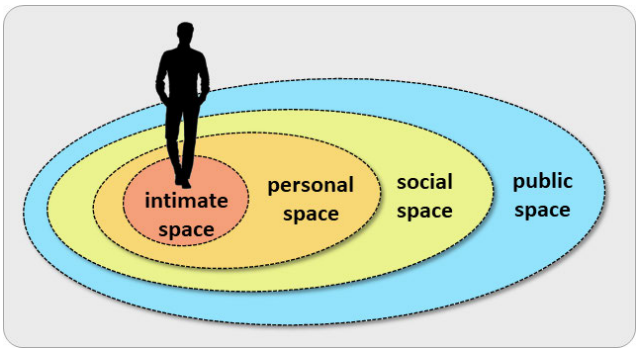


FIGURE 2. A graphical representation of the personal spaces that are used in proxemics.

Similarly, *Attachment Theory*, probably the development model most widely accepted in child psychiatry, revolves around the ability of children and parents to establish and maintain physical proximity [13]. Finally, the different modulation of interpersonal distances is known to be one of the main obstacles in intercultural communication [37].

The above suggests that dealing with interpersonal distances means to deal with evolutionary, developmental and cultural forces that shape, to a significant extent, our everyday life. As a consequence, the role of technologies for the analysis of such distances becomes crucial during pandemics, given that they must mediate between the forces above, responsible for the human tendency to get too close to avoid contagion, and the pressure of prophylactic measures, artificially designed to fight a pathogen inaccessible to our senses and cognition.

One possible solution is to go beyond simply measuring how far we are from one another, as most of the applications on the market are doing (see Sec. II-C) and try to make sense of what distances mean. In other words, it is necessary to inform technologies with principles and laws of *Proxemics*, the psychology area showing how people convey social meaning through interpersonal distances and, ultimately, how social and physical dimensions of space interplay with one another [74].

Proxemics is strictly linked to the definition of people gatherings, namely groups, and as such, it depends on its spatial organization and the number of people involved. In general, the surrounding space around a person is characterized by *interpersonal distance classes* [38], namely: intimate, personal, peri-personal or social, and public spaces (see Fig. 2), all associated to different social distances, in turn, also dependent by the degree of kinship and familiarity between the subjects and by the geometrical configuration and size of the environment in which an interplay occurs. A blind application of SD rules, encouraging to stay further than 1-2 meters, will eliminate an entire interpersonal distance class and all of the social interactions which take play within it, including for example those between children and relatives. Indeed, as can be noticed, behavior, social interactions, and space arrangements are tightly coupled, and affect each other. This is why it is important to take into consideration all these aspects when constraints in this respect are to be imposed, in particular when people health is in play.

For all these reasons, the focus of this paper lies on *Visual Social Distancing* (VSD), *i.e.* on approaches relying on video cameras and other imaging sensors (see Fig. 1 for an example) to analyse the proxemic behaviour of people. The main reason behind the choice of VSD is that computer vision and social signal processing have already developed methods for automatic measurement and understanding of interpersonal distances (see Sec. II for more details). Furthermore, VSD approaches have shown advantages that can complement other technologies like, *e.g.*, mobile applications based on large-scale mobility patterns. In particular, VSD approaches can characterize interpersonal distances in terms of social relations (*e.g.*, whether people at a certain distance are friends, family members or partners), thus allowing one to modulate interventions according to such an information. Furthermore, vision-based technologies can detect contextual information helpful to understand whether social distancing rules are actually being broken or not. For example, VSD can understand whether people get too close because the situation makes it necessary (*e.g.*, when someone rescues a person in trouble) or whether the distance is not a problem (*e.g.*, when people wear personal protective equipment and can safely stay close).

Finally, VSD helps to understand the reason why some people stand close, distinguishing whether they are socializing among themselves, or if they are interacting with the environment (as, for example, looking at a timetable in the airport), thus suggesting the most proper countermeasures to ensure SD (*e.g.*, rising an audio alarm to discourage social interactions or putting markers into the floor so that people can watch the time table while keeping the right distances).

The advantages of VSD appears to be of particular importance since at the moment social distancing rules have to be expressed in simplistic terms (*e.g.*, people have to be at least 2 meters far from one another) requiring one to

distinguish between the intention (avoid contagion) and the rule (keep a minimum interpersonal distance). Such a distinction, evident to humans, poses a real and new challenge to a computational algorithm for VSD that could solve the problem by leveraging, for instance, the use of contextual information. Differently, the number of false alarms would be so high that any benefit resulting from the use of technology would be canceled.

In the following, we will discuss in detail the VSD problem and its connection to the Computer Vision and Social Signal Processing research domains. Starting from a geometrical point of view, *i.e*. estimating inter-personal distances between people from an image, we show that this first step does not take into account scene and social context. For this reason, a further stage needs to elaborate the geometrical VSD in order to interpret whether the violation of the distance is a real cause of alert or an acceptable situation (*e.g*., a family walking). Then, we contextualise the VSD in different application domains and we finally conclude with a description of the possible ethical shortcomings of VSD.

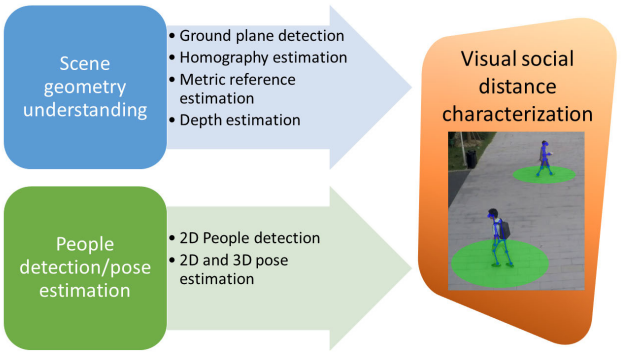


FIGURE 3. The VSD problem requires the solution of different problems. The estimation of a local metric reference system using the scene geometry (blue box) and the detection of pedestrians and (possibly) their pose in the image (green box). This information provides a geometrical measure of interpersonal distance that has to be interpreted given the social context of the scene (orange box).

II. VISUAL SOCIAL DISTANCE ESTIMATION

Estimating the VSD requires one to solve a few classical Computer Vision and Social Signal Processing tasks as identified in Fig. 3, namely, scene geometry understanding (Sec. II-A), person detection/body pose estimation (Sec. II-B) and social distance characterization (Sec. II-C). Indeed, the geometry of the scene is important to define a local reference system for measuring inter-personal distances. Clearly, a second and important task is the detection of people in the scene in possibly crowded environments. Once the target people are correctly localised in a scene, their distance can be locally estimated in order to realize if the mutual distance is lower than a threshold (*e.g*. 1m or 2m). Afterwards, this metric information is analysed to output whether there is a violation of the protocol or the short distance is due to a legitimate situation, *e.g*., a family walking together.

In the following, we will describe these modules in detail re-targeting, when possible, previous Computer Vision methods that can provide a solution to these problems.

A. SCENE GEOMETRY UNDERSTANDING

The task of measuring social distancing from images requires the definition of a (local) metric reference system. This problem is strongly related to the single view metrology topic [20] as we consider the most common case of a fixed camera. An initial solution for estimating inter-personal distances requires the identification of the ground plane where people walk [1], [40], [50], [62], [79], [102], [107]. Such ground plane serves in many video-surveillance systems to visualise the scene as a bird's eye view for ease of visualisation and data statistics representation. Many works impose the assumption that the ground plane is planar. Then, the problem is to estimate a homography given some reference elements (*e.g*., known objects or manual measurements) extracted from the scene or using the information of detected vanishing points in the image [4], [5], [20], [51], [58], [60], [69], [77], [109], [113]. Another common approach is to calibrate fixed cameras by observing the motion of dynamic objects such as pedestrians [53], [59], [91], [95]. Recently, approaches based on deep learning attempt at estimating directly camera pose and intrinsic parameters on a single image [41], [57].

Even if these approaches might provide an estimate of the camera intrinsic/extrinsic parameters and the detection of the ground plane, still VSD estimation requires a metric reference. Such an information can be coarsely computed in the scene given objects of known dimension or by using a standardised height of pedestrians as a rule of thumb [8], [100], [102]. Given the current state of the art, we have the following observations related to the geometrical aspects of VSD:

- Although the planarity constraint might not hold for the entire image, VSD has to do a local estimation of proximity for which is safe to relax the scene being piece-wise planar.
- Self-calibration approaches highly rely on the existence of a Manhattan world (*e.g*. vanishing points are detectable) or pedestrian walking in straight trajectories, which limit the applicability of such methods. Estimating depth from single image might be a viable option, but a metric reference is still needed.
- Estimating a metric reference for precise SD measures from images is an issue. Such reference extracted from pedestrians might be unreliable given the variations in anthropometric characteristics. Reasoning on the geometrical context of the scene (*e.g*., object shapes) can lead to a more robust metric estimate.

It is also important to emphasizes here that VSD is a simpler problem than estimating every metric distances among people in any position in the image. Estimating social distance is necessary when two or more pedestrians get close enough for triggering the necessity of a measure. At this point, a local reference system can be estimated and metric

references can be leveraged by using surrounding objects and the height of the local cluster of people.

To this end, Social Signal Processing (SSP) findings related to the detection of the group formation and tracking [6], [21], [28], [84], [88] can be useful to identify which pedestrians should be selected for estimating the local VSD. These local estimates with an associated metric reference can be useful whenever a global camera pose is hard to estimate or if the single ground plane assumption is violated, a likely occurrence in an unconstrained scenario.

B. PERSON DETECTION AND POSE ESTIMATION

Person detection has reached impressive performance in the last decade given the interest in automotive industry and other application fields [9]. Real-time approaches can now estimate people pose even in complex scenarios [14] and even reconstruct the 3D mesh of the person body [35]. The majority of the approaches estimate not only people location as a bounding box but also 2D stick-like figures, so conveying a schematic representation of the pose. Recently, several methods augment 2D poses in 3D or infer directly a 3D pose in a normalised reference system [11], [63], [67], [68], [71], [75], [93], [103], [114].

Capturing diffused small SDs with Computer Vision requires to localize multiple people, realizing the hardest scenario for pedestrian detection techniques. Specific pedestrian detection techniques have been designed to work in crowded scenes [29], [55], [97], [106], where saliency-based masks are often preferred to skeleton-based representations. When the image resolution becomes too low to spot single people, regression-based approaches are employed [12], [15], [54], [80], [92], [104], [111], providing in some case density measures [73], [86], [87], [110]. This information, merged with a geometric model of the scene, will directly lead to a measure of the average SD in the field of view. Obviously, regression or density-based approaches cannot provide additional cues on pose which are highly important for capturing human actions and interactions. To fill this gap, ad-hoc approaches individuate general crowd activities, classifying them as normal or not (e.g. a person collapsing and many people getting close) [25], [34], [70], [72].

Recently, new efforts address human detection and body pose estimation in crowded environments [32], [52], the very same scenario social distancing is dealing with. Yet, finding the location of people in such cases is necessary for alerting or creating statistics of overcrowded areas. To this end, a people detection module has to be robust to severe self and other objects/people occlusions, different image scales, and indoor/outdoor scenarios. Although a person detection (*i.e.*, without the pose) may be enough for estimating the VSD, finding joints and body parts of pedestrian has certain advantages. This is due the fact that to obtain an approximate metric reference, or even calibrating cameras, usually the person height is used as a coarse proxy as computed from a bounding box or by more precise techniques [8], [24], [36], [100], [102]. However, bounding boxes do not account for different body poses (*e.g.*, sitting, riding) that might negatively impact the estimate of height and thus a wrong VSD. Another issue is related to occlusions, *i.e.* how reliable is to extract a person height without having a full body information? This is necessary in the likely case of crowded environments or whenever an object partially hides person body parts (*e.g.*, a person seated at a desk).

Given a metric reference from scene geometry and the position/pose of the people in the scene, the SD can be calculated as a distance on the ground plane (feet/body pose centroid) among all the possible detected pedestrians. As previously discussed, this information can be estimated locally or pairwise in order to reduce the complexity of estimating a global reference system for the whole image.

C. VISUAL SOCIAL DISTANCE CHARACTERIZATION

Social distances should be complemented with additional contextual information to understand whether social distancing rules are actually being broken or not, consequently suggesting the most proper reaction.

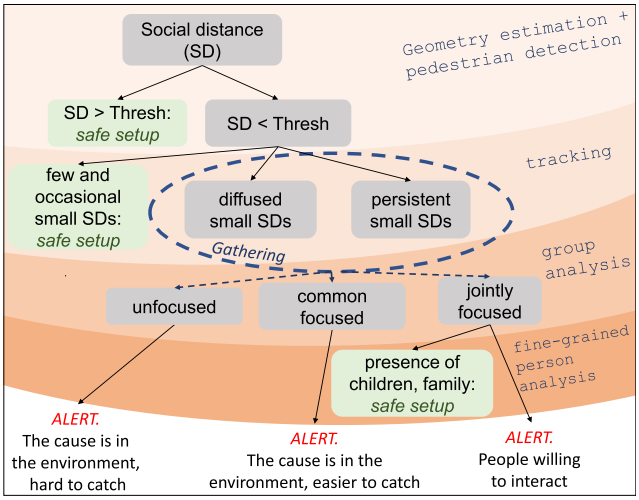


FIGURE 4. How to characterize social distances. Together with the taxonomy explained in the text, we specify in courier new the Computer Vision technologies to access a particular level of SD specification. The more detailed the SD characterization, the more advanced yet fragile Computer Vision technology is.

Fig. 4 reports a multi-layer pipeline, which will be detailed in the following, indicating which information can be accessed with the current Computer Vision technology. The deeper the layer (indicated by a darker color), the finer the visual analysis which is needed and the harder the corresponding request for Computer Vision.

As previously stated, SDs taking values above a certain threshold would certainly comply with social distancing rules. On the contrary, the presence of SDs under a certain threshold ($SD < Thresh$ in Fig. 4) *could* be considered as breaking the rules, but actually many are the scenarios where this should not raise any concern.

For example, occasional small SDs holding for few frames, especially in a crowded scenario (the *few and occasional*

small SDs blob in the figure), can be allowed, considering that they are detected by automatic approaches which are typically not so accurate. Instead, small SDs can be critical if they are 1) *diffused* and/or 2) *persistent*.

In the former case, a high percentage of small SDs is characterizing the monitored area: This may occur at a crossing intersection or walking in a corridor of an airport. Here, many persons stand close, without an explicit will, possibly for a short (∼seconds) period. Computer Vision helps here providing robust approaches for pedestrian detection and counting (as discussed in Sec. II-B).

Persisting small SDs mean that specific people stand close to each other for a certain time interval. This condition addresses a more interesting situation, since it likely indicates people staying close *by intention*. Under a Computer Vision point of view, persisting small SDs are more difficult to capture, since they require people to be tracked continuously, maintaining their identity. Interested readers may refer to [23], [56], discussing the problem of tracking in crowded situations.

In Social Signal Processing, diffused and/or persisting small SDs individuate *gatherings* [30], [31], [47], [84], which are generically addressed as ''groups'' or ''crowds'' in Computer Vision. The term gathering refers precisely to ''any set of two or more individuals in mutual presence at a given moment who are having some form of social interaction'' [31]. With the expression *social interaction* we mean the process by which we act and react to those around us [84]. Many types of gatherings are documented in the sociological literature, depending on:

- number of people being part of the gathering;
- *type* of social interaction;
- spatial dynamics.

As for the number of people, we may have small (2 to 6 people), medium (7 to 12-30 people), or large gatherings (larger than 13-31) [39].

Small gatherings occur in private (home, private garden, car), semi public (classroom, office, club, party area), and public places (open plaza, transportation station, walkway, park, street). Medium gatherings may occur in private, but mostly in semi-public and public places, the latter being also the preferred venues of large gatherings [39].

As for the type of social interaction, *unfocused interaction* occurs whenever individuals find themselves by circumstance in the immediate presence of others. For instance, when forming a queue, or when walking in the crowded corridor of an airport. On such occasions, simply by virtue of the reciprocal presence, some form of interpersonal communication must take place regardless of the individual intent.

For our study, having people forming an unfocused gathering and exhibiting small SD may indicate a problematic scenario, since it is the *context* which encourages the formation of tight gatherings and not the will of people. As a consequence, to avoid such type of gathering may require a change of the context itself, for example discouraging the queues with markers on the floor, or creating lanes with barricades.

Conversely, a *focused interaction* occurs whenever two or more individuals willingly agree – although such an agreement is rarely verbalised – to sustain for a period a single focus of cognitive and visual attention [30].

Focused gatherings can be further distinguished in *common focused* and *jointly focused* [46]. In the former case, the focus of attention is common and not reciprocal, for example watching a timetable screen at the airport, watching a map in the metro station, being at a concert. Common focused gathering exhibiting small SDs can be dealt more easily than in presence of unfocused gathering, since in this case the reason of the gathering is easier to be captured, which is the item or event attracting the common attention of people.

Jointly focused gathering, finally, entails the sense of mutual, instead of merely common, activity. In this case, the participation is not at all peripheral but engaged: people are – and display to be – mutually involved [31]. Since the presence of a jointly focused gathering depend on the will of people, when this is characterized by a small social distance, it can be discouraged by simply alerting the people about the ongoing critical setup. An exception for this scenario occurs when a jointly focused gathering involves children, elderly requiring care, or anybody with an impairment that have to be accompanied, and are usually at physical contact with their relatives or caregivers.

Some combinations of these attributes give rise to specific types of gatherings (shown in Fig. 5), some of them addressed by explicit definitions: small gatherings of jointly focused people, mostly static, are dubbed by Kendon *free-standing conversational groups* [47], highlighting their spontaneous aggregation/disgregation nature, implying that their members are jointly focused, and specifying their mainly-static proxemic layout. Large gatherings of unfocused people are named *casual crowds* [10], commonly focused large gathering refers to *spectator crowd* [10] and, finally, large gatherings of jointly focused people are *demonstration/protest* or *Acting crowd* [66].

As anticipated above, most Computer Vision approaches do not build upon this taxonomy, distinguishing merely gatherings depending on the number of individuals involved, leading to groups (= small gathering for sociology) and crowds (= large gatherings), with some exception presented in the following. Groups have been usually identified exploiting positional and velocity cues (people in a group are close and move with similar oriented velocity) [33], [42], [76], [78], [81], [82], [88]–[90], [94]. Explicit focus on free-standing conversational groups is given in [21], [44], [83], [84], [98], [99]. In most of these latter approaches, positional and velocity cues are enriched by pose information, fully capturing the people proxemics. Coming back to the characterization of SDs, and to Fig. 4, joint-focused groups where people stand closer than a given threshold requires maximal attention, since their vicinity is by choice, and not by external circumstances. At this level of

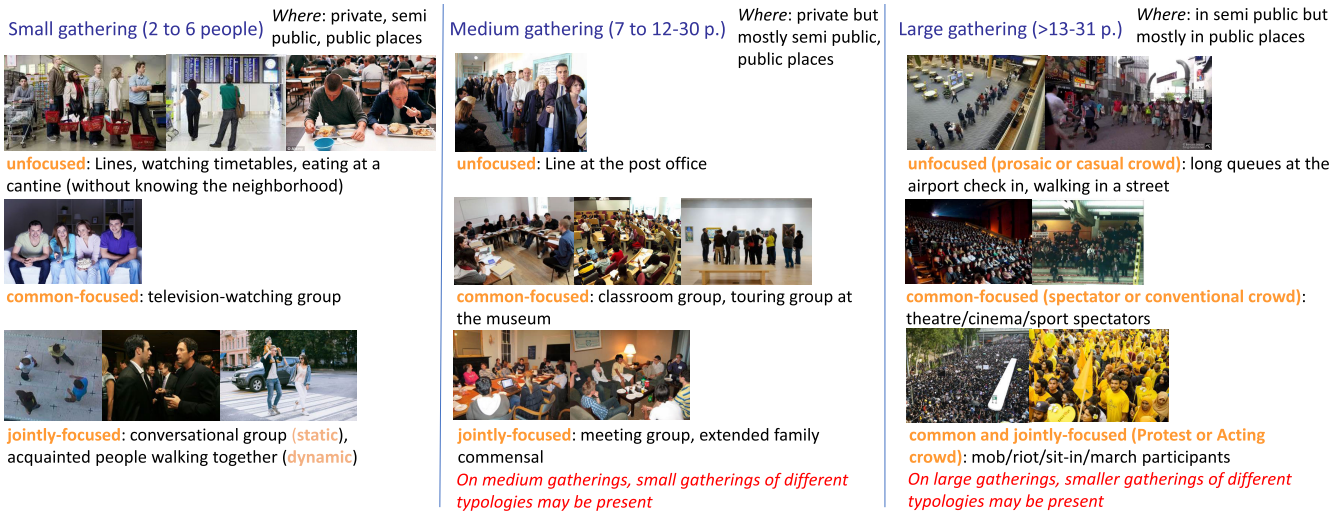


FIGURE 5. Different typologies of gathering, depending on the number of individuals involved and type of social interaction. (cfr. https://vips.sci.univr.it/research/fformation/).

characterization, avoiding false alarms would mean to focus on the age of the interactants: Having children in a small gathering would probably indicate a family. Approaches like [112] estimate the age from pedestrian detections, but solving this task efficiently seems to be still at its early stages.

As for the modelling of crowd, some approaches allow to estimate the number of individuals [12], [15], [54], [80], [104] or their density [86], [87], [110]. Social Signal Processing approaches for large gatherings focus specifically on common-focused formations (i.e. *spectator crowd*), still capturing proxemic cues including the body pose [18], [19]. Medium gatherings have never been properly addressed neither by Computer Vision nor Social Signal Processing literature.

Finally, we should consider the static/dynamic axis concerning the degree of freedom and flexibility of the spatial, positional, and orientational organisation of gatherings. Distinguishing between uncommon, common-focused and jointly focused is hard, since when a gathering is moving, their members spend attention to follow safe trajectories, avoiding collisions. Therefore, the aforementioned taxonomy holds especially for static formations, with few exceptions [17]. When people are moving, the only valid distinction between Computer Vision and Social Signal Processing is related to small and large gatherings.

For small gatherings, temporal information allows one to provide stronger grouping estimations, analyzing pedestrians motion paths instead of static positions [64], [88], [101]. For large gathering, many approaches identify dominant flows and segment crowd according to coherent motion [16], [45], [96], [105], and to identity collective/abnormal behaviors [25], [34], [70], [72].

Summarizing, once small SDs are detected, it is necessary to understand if they are persistent and/or diffused in the scene. Then, proxemic analysis is needed to characterize the gatherings which are generating those SDs. Unfocused gatherings indicate that SDs are caused by no explicit will, while common-focused gatherings usually occur because of the presence of precise environmental conditions (a specific item/landmark of interest or an event attracting the attention). Jointly-focused gatherings indicate explicit will of interacting, and could be further described by capturing the age of interactants (kinship). Each of these formations may demand for different interventions, thus going beyond the simple alarm when SDs are too small, while diminishing false positive alarms.

Computer vision approaches following this taxonomy exist for jointly focused (small) gatherings, (large) common focused gatherings, and show that positional and body pose cues are of primary importance. Further work has to be done to cover all the possible types of gatherings, as current technology is still struggling to achieve a solution, especially when they are composed by several people.

III. BEYOND SOCIAL DISTANCING: APPLICATIONS

While potentially playing a crucial role in the case of a virus outbreak, technology developed for the analysis of social distancing can be useful in a large number of application domains that, therefore, can benefit from the approaches proposed in this work.

The detection of mental health issues is one of the areas that will benefit most from the application of AI,[1] supported by the World Health Organization observing that a pathology like depression affects around 300 millions people around the world [108]. In such a particular case, the tendency to avoid physical proximity and engagement with others is an important symptom. The technologies proposed in this work

[1]According to the *Gartner Group*, a relevant strategic consulting company: http://www.gartner.com/smarterwithgartner/ 13-surprising-uses-for-emotion-ai-technology/

can also help to monitor the increase of SD, especially when it is hard to observe. Similarly, the analysis of interpersonal distances can help to identify children with insecure attachment, known to manifest their condition through irregular proximity patterns (among other cues) [13].

Another important domain where the analysis of SD is important is social robotics. In particular, the *International Federation of Robotics* pointed out that public relation robots are the fastest growing area of service robotics with estimates in sales moved from a total of USD 319 million in the period 2015-2017, to a total of USD 746 million between 2018 and 2020. In this field, the use of proxemics appears to be particularly important to ensure that a robot is perceived to play correctly its role (*e.g.*, whether it is expected to be a servant or a companion in playing) [49], and to establish a sense of intimacy [48], an aspect of focal importance in assistive robotics. In addition, distance plays a major role along one of the five Godspeed dimensions typically used to assess the quality of human-robot interaction, namely perceived safety [3].

In the last years, most major companies have introduced training to avoid unconscious bias, *i.e.* the tendency to discriminate certain categories of people without being aware of it. This happens not only for ethical reasons, but also because *McKinsey* has shown that companies ensuring diversity in their workforce, especially at the top management levels, are 30% more likely to be above national median in terms of financial returns [43]. As a consequence, major companies like Facebook (https://managingbias.fb.com) and Google (https://diversity.google) adopt implicit training programs. Furthermore, Forbes estimates that the market of implicit bias and diversity training has reached a value close to USD 9 billion-a-year (http://goo.gl/R53xn4). Unconscious bias leaves different traces in nonverbal behaviour and one of these is the increase of physical distances (*e.g.*, see [65]). Therefore, automatic technologies for proxemic analysis can help to detect the phenomenon, contributing to protect the potential victims, and train the bias bearers to identify and attenuate their tendencies to discriminate others.

A large number of studies show that the architectural design of space influences the behaviour of its inhabitants [61]. For example, a simple line on the floor separating right and left side of a corridor makes the flow of people through it more ordered [2]. Similarly, the restructuring of Westminster in the UK aims at improving the efficiency of parliamentary works, but encounters the opposition of Parliament workers afraid of disrupting established traditions by the change of the way space is organised [85]. Until now, the study of these phenomena has been performed mainly through ethnographic observations, but the development of technologies for proxemic analysis can certainly help by producing more objective and quantitative data about the change in habits of the people. This is in line with previous works about the study of organisations through the use of smart badges detecting who is in proximity with whom in an organisation [27].

Besides the application scenarios above, likely to benefit from the technologies presented in this work in the future, there are established domains that can benefit from models of mutual distancing. For example, Augmented and Mixed Reality technologies can provide more immersive and engaging experiences through the inclusion of virtual characters capable to move with respect to users like humans do with respect to one another. Similarly, surveillance systems can further refine their ability to detect events of interest in a given environment like, e.g., an aggression in a public space. Finally, technologies analysing interpersonal distances can be of help to social psychologists that investigate the dynamics of social interactions. In other words, far from being exhaustive, the list of application domains listed in this section still provides an indication of how wide the application of VSD can be once the COVID-19 outbreak, at the origin of the most recent interest towards interpersonal distances, will be over.

IV. PRIVACY AND ACCEPTABILITY CONCERNS

Optical cameras are the most widespread sensors for VSD measurements and the acceptance of this monitoring technology can be difficult since it clearly raises privacy concerns. Video footage may disclose the identity of the persons captured and in general recording is regulated by strict laws, both at national and international level. Moreover, potential attacks to the video transmission channels and to storage servers can pose a relevant security issue.

However, the current computer vision technology is now mature to manage effectively privacy concerns. Alternatives benefit from the usage of the so-called *smart* cameras [7] which, having computing capability onboard, are able to process video data up to a certain capacity. By adopting a *privacy-by-design* principle, a first option is to process video sequences internally, while measuring and transfer only VSD estimates without any visual data, thus sensible information, being transmitted to the remote control operative room. This is of crucial importance for VSD, since as we have been shown in the previous sections, accurate estimates may require the identification of kinship. This sensible information is clearly not necessary to be disclosed for estimating VSD, and any possible leak has to be avoided.

At the same time it is worth noting that VSD technology exhibits features that differentiate it from other apparently safer alternatives, as geolocation data collected from mobile applications. VSD techniques are in fact non-invasive and mostly non-collaborative, meaning that the user does not need to provide ID personal data. Tracing technologies, on the contrary, need to be fed with sensible data and even when this is totally anonymized, recent research [22] proves that individuals may still be identified by a few information – four spatio-temporal points allows one to uniquely identify 95% of people in a mobile phone database of 1.5 million subjects, and 90% of people in a credit card database of 1 million individuals.

V. CONCLUSIONS

In this paper, we have presented the VSD problem as the estimation and characterization of inter-personal distances from images. Solving such problems allows a quick screening of the population for detecting potential behaviours that can cause a health risk, especially related to recent pandemic outbreaks. We pointed out that VSD is not only a Computer Vision problem related to geometrical proxemic since people distancing has to be weighted given the social context in the current scene. Close relationships can allow closer interpersonal distances as well as being a caregiver of individuals with fragile conditions. We have shown that understanding such social context is a compelling problem in the literature of signal social processing that requires further research efforts for a reliable solution. As the solution is intertwined with the decoding of social relationships from images, there are strong ethical and privacy concerns that need to be addressed with novel privacy-by-design solutions. Past this grievous global crisis, VSD has still an important role in several application fields thus providing a continuous source of interest in this new problem.

ACKNOWLEDGMENT

The authors would like to thank P. Morerio, M. Aghaei, G.L. Bailo and M. Bustreo for the results on social distancing estimation in the figures.

REFERENCES

[1] S. A. Abbas and A. Zisserman, "A geometric approach to obtain a Bird's eye view from an image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4095–4104.

[2] P. Ball, *Critical Mass*. New York, NY, USA: Arrow Books, 2004.

[3] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Social Robot.*, vol. 1, no. 1, pp. 71–81, Jan. 2009.

[4] J.-C. Bazin and M. Pollefeys, "3-line RANSAC for orthogonal vanishing point detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 4282–4287.

[5] J.-C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys, "Globally optimal line clustering and vanishing point estimation in manhattan world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 638–645.

[6] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Syst.*, vol. 30, no. 2, pp. 115–127, May 2013.

[7] A. N. Belbachir, *Smart Cameras*, vol. 2. New York, NY, USA: Springer, 2010.

[8] C. BenAbdelkader and Y. Yacoob, "Statistical body height estimation from a single image," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–7.

[9] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014, pp. 613–627.

[10] H. Blumer, *Collective behavior*. Indianapolis, IN, USA: Bobbs-Merrill Company Incorporated, 1957.

[11] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 561–578.

[12] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. ACM Multimedia Conf. (MM)*. New York, NY, USA: ACM, 2016, pp. 640–644.

[13] J. Bowlby, *Attachment and Loss*. London, U.K.: The Hogarth Press and the Institute of Psycho-Analysis, 1969.

[14] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 17, 2019, doi: 10.1109/TPAMI.2019.2929257.

[15] A. B. Chan, Z.-S. John Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[16] A. M. Cheriyadat and R. J. Radke, "Detecting dominant motions in dense crowds," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 4, pp. 568–581, Aug. 2008.

[17] T. M. Ciolek and A. Kendon, "Environment and the spatial arrangement of conversational encounters," *Sociol. Inquiry*, vol. 50, nos. 3–4, pp. 237–271, Jul. 1980.

[18] D. Conigliaro, P. Rota, F. Setti, C. Bassetti, N. Conci, N. Sebe, and M. Cristani, "The S-HOCK dataset: Analyzing crowds at the stadium," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2039–2047.

[19] D. Conigliaro, F. Setti, C. Bassetti, R. Ferrario, and M. Cristani, "Viewing the viewers: A novel challenge for automated crowd analysis," in *Proc. Int. Conf. Image Anal. Process. (ICIAP)*. Berlin, Germany: Springer, 2013, pp. 517–526.

[20] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 123–148, 2000.

[21] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. D. Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of F-formations," in *Proc. Brit. Mach. Vis. Conf.* BMVA, 2011, p. 4.

[22] Y.-A. de Montjoye *et al.*, "On the privacy-conscientious use of mobile phone data," *Sci. Data*, vol. 5, no. 1, pp. 1–6, Dec. 2018.

[23] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixe, "CVPR19 tracking and detection challenge: How crowded can it get?" 2019, *arXiv:1906.04567*. [Online]. Available: http://arxiv.org/abs/1906.04567

[24] R. Dey, M. Nangia, K. W. Ross, and Y. Liu, "Estimating heights from photo collections: A data-driven approach," in *Proc. 2nd Ed. ACM Conf. Online Social Netw. (COSN)*. New York, NY, USA: ACM, 2014, pp. 227–238.

[25] C. Direkoglu, M. Sah, and N. E. O'Connor, "Abnormal crowd behavior detection using novel optical flow-based features," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[26] R. Dunbar, *Human Evolution*. Baltimore, MD, USA: Penguin, 2014.

[27] N. Eagle and K. Greene, *Reality Mining: Using Big Data to Engineer a Better World*. Cambridge, MA, USA: MIT Press, 2014.

[28] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012.

[29] Z. Ge, Z. Jie, X. Huang, R. Xu, and O. Yoshie, "PS-RCNN: Detecting secondary human instances in a crowd via primary object suppression," 2020, *arXiv:2003.07080*. [Online]. Available: http://arxiv.org/abs/2003.07080

[30] E. Goffman, *Encounters; Two Studies in the Sociology of Interaction*. Indianapolis, IN, USA: Bobbs-Merrill, 1961.

[31] E. Goffman, *Behavior in Public Places: Notes on the Social Organization of Gatherings*. New York, NY, USA: Free Press, 1966.

[32] T. Golda, T. Kalb, A. Schumann, and J. Beyerer, "Human pose estimation for real-world crowded scenarios," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[33] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz, "Detecting social situations from interaction geometry," in *Proc. IEEE 2nd Int. Conf. Social Comput.*, Aug. 2010, pp. 1–8.

[34] X. Gu, J. Cui, and Q. Zhu, "Abnormal crowd behavior detection by using the particle entropy," *Optik*, vol. 125, no. 14, pp. 3428–3433, Jul. 2014.

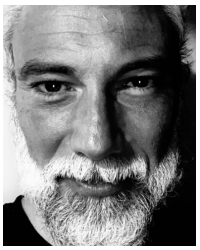
[35] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[36] S. Gunel, H. Rhodin, and P. Fua, "What face and body shapes can tell us about height," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1819–1827.

[37] E. T. Hall, *The Silent Language*. New York, NY, USA: Anchor Books, 1959.

[38] E. T. Hall, *The Hidden Dimension*. New York, NY, USA: Doubleday & Co, 1966.
[39] A. P. Hare, "Group size," *Amer. Behav. Scientist*, vol. 24, no. 5, pp. 695–708, 1981.
[40] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 151–172, Jul. 2007.
[41] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambaretto, S. Hadap, and J.-F. Lalonde, "A perceptual measure for deep single image camera calibration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2354–2363.
[42] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in *Proc. 13th Int. Conf. Multimodal Interface (ICMI)*, Nov. 2011, pp. 231–238.
[43] V. Hunt, D. Layton, and S. Prince, "Why diversity matters," McKinsey, Shanghai, China, Tech. Rep., 2015.
[44] S. Inaba and Y. Aoki, "Conversational group detection based on social context using graph clustering algorithm," in *Proc. 12th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2016, pp. 526–531.
[45] K. Kang and X. Wang, "Fully convolutional neural networks for crowd segmentation," 2014, *arXiv:1411.4464*. [Online]. Available: http://arxiv.org/abs/1411.4464
[46] A. Kendon, "Goffman's approach to face-to-face interaction," in *Erving Goffman: Exploring the Interaction Order*, P. Drew and A. Wootton, Eds. Hoboken, NJ, USA: Wiley, 1988, pp. 14–40.
[47] A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[48] J. Kennedy, P. Baxter, and T. Belpaeme, "Nonverbal immediacy as a characterisation of social behaviour for human–robot interaction," *Int. J. Social Robot.*, vol. 9, no. 1, pp. 109–128, Jan. 2017.
[49] K. L. Koay, D. S. Syrdal, M. Ashgari-Oskoei, M. L. Walters, and K. Dautenhahn, "Social roles and baseline proxemic preferences for a domestic service robot," *Int. J. Social Robot.*, vol. 6, no. 4, pp. 469–488, Nov. 2014.
[50] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: Establishing a common coordinate frame," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 758–767, Aug. 2000.
[51] B. Li, K. Peng, X. Ying, and H. Zha, "Vanishing point detection using cascaded 1D Hough transform from single images," *Pattern Recognit. Lett.*, vol. 33, no. 1, pp. 1–8, Jan. 2012.
[52] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10863–10872.
[53] J. Liu, R. T. Collins, and Y. Liu, "Surveillance camera autocalibration based on pedestrian height distributions," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*. BMVA, 2011, p. 144.
[54] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "DecideNet: Counting varying density crowds through attention guided detection and density estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5197–5206.
[55] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6459–6468.
[56] W. Liu, M. Salzmann, and P. Fua, "Estimating people flows to better count them in crowded scenes," 2019, *arXiv:1911.10782*. [Online]. Available: http://arxiv.org/abs/1911.10782
[57] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro, "Deep single image camera calibration with radial distortion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11817–11825.
[58] X. Lu, J. Yaoy, H. Li, Y. Liu, and X. Zhang, "2-line exhaustive searching for real-time vanishing point estimation in manhattan world," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 345–353.
[59] F. Lv, T. Zhao, and R. Nevatia, "Camera calibration from video of a walking human," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1513–1518, Sep. 2006.
[60] M. J. Magee and J. K. Aggarwal, "Determining vanishing points from perspective images," *Comput. Vis., Graph., Image Process.*, vol. 26, no. 2, pp. 256–267, May 1984.
[61] H. Mallgrave, *Architecture and Embodiment: The Implications of the New Sciences and Humanities for Design*. Evanston, IL, USA: Routledge, 2013.
[62] Y. Man, X. Weng, X. Li, and K. Kitani, "GroundNet: Monocular ground plane normal estimation with geometric consistency," in *Proc. 27th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2019, pp. 2170–2178.
[63] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2640–2649.
[64] R. Mazzon, F. Poiesi, and A. Cavallaro, "Detection and tracking of groups in crowd," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 202–207.
[65] C. McCall, J. Blascovich, A. Young, and S. Persky, "Proxemic behaviors as predictors of aggression towards black (but not White) males in an immersive virtual environment," *Social Influence*, vol. 4, no. 2, pp. 138–154, Apr. 2009.
[66] C. McPhail and R. T. Wohlstein, "Individual and collective behaviors within gatherings, demonstrations, and riots," *Annu. Rev. Sociol.*, vol. 9, no. 1, pp. 579–600, Aug. 1983.
[67] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "XNect: Real-time multi-person 3D motion capture with a single RGB camera," 2019, *arXiv:1907.00837*. [Online]. Available: http://arxiv.org/abs/1907.00837
[68] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.
[69] F. M. Mirzaei and S. I. Roumeliotis, "Optimal estimation of vanishing points in a manhattan world," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2454–2461.
[70] S. Mohammadi, F. Setti, A. Perina, M. Cristani, and V. Murino, "Groups and crowds: Behaviour analysis of people aggregations," in *Proc. Int. Joint Conf. Comput. Vis., Imag. Comput. Graph.* Cham, Switzerland: Springer, 2016.
[71] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2823–2832.
[72] A. Pennisi, D. D. Bloisi, and L. Iocchi, "Online real-time crowd behavior detection in video sequences," *Comput. Vis. Image Understand.*, vol. 144, pp. 166–176, Mar. 2016.
[73] A. Rangel-Huerta, A. L. Ballinas-Hernández, and A. Muñoz-Meléndez, "An entropy model to measure heterogeneity of pedestrian crowds using self-propelled agents," *Phys. A, Stat. Mech. Appl.*, vol. 473, pp. 213–224, May 2017.
[74] V. Richmond and J. McCroskey, *Nonverbal Communication in Interpersonal Relations*. London, U.K.: Pearson, 1999.
[75] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, May 2020.
[76] P. Rota, N. Conci, and N. Sebe, "Real time detection of social interactions in surveillance video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2012.
[77] C. Rother, "A new approach to vanishing point detection in architectural environments," *Image Vis. Comput.*, vol. 20, nos. 9–10, pp. 647–655, Aug. 2002.
[78] N. Sanghvi, R. Yonetani, and K. M. Kitani, "MGpi: A computational model of multiagent group perception and interaction," in *Proc. Int. Conf. Auto. Agents Multiagent Syst. (AAMAS)*, 2019, pp. 1196–1205.
[79] S. Se and M. Brady, "Ground plane estimation, error analysis and applications," *Robot. Auto. Syst.*, vol. 39, no. 2, pp. 59–71, May 2002.
[80] F. Setti, D. Conigliaro, M. Tobanelli, and M. Cristani, "Count on me: Learning to count on a single image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1798–1806, Aug. 2018.
[81] F. Setti and M. Cristani, "Evaluating the group detection performance: The GRODE metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 566–580, Mar. 2019.
[82] F. Setti, H. Hung, and M. Cristani, "Group detection in still images by F-formation modeling: A comparative study," in *Proc. 14th Int. Workshop Image Anal. for Multimedia Interact. Services (WIAMIS)*, Jul. 2013, pp. 1–4.
[83] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-scale F-formation discovery for group detection," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3547–3551.
[84] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PLoS ONE*, vol. 10, no. 5, May 2015, Art. no. e0123783.

[85] S. Siebert, "Two visions of the future: Restoration and renewal of the UK Parliament," in *Proc. Annu. Meeting Soc. Advancement Socio-Econ.*, 2019.
[86] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
[87] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, May 2018.
[88] F. Solera, S. Calderara, and R. Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 995–1008, May 2016.
[89] M. Swofford, J. Peruzzi, and M. Vázquez, "Conversational group detection with deep convolutional networks," Oct. 2018, *arXiv:1810.04039*. [Online]. Available: https://arxiv.org/abs/1810.04039
[90] M. Swofford, J. Peruzzi, N. Tsoi, S. Thompson, R. Martín-Martín, S. Savarese, and M. Vázquez, "Improving social awareness through DANTE: Deep affinity network for clustering conversational interactants," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW1, pp. 1–23, May 2020.
[91] Z. Tang, Y.-S. Lin, K.-H. Lee, J.-N. Hwang, and J.-H. Chuang, "ESTHER: Joint camera self-calibration and automatic radial distortion correction from tracking of walking humans," *IEEE Access*, vol. 7, pp. 10754–10766, 2019.
[92] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, "PaDNet: Pan-density crowd counting," *IEEE Trans. Image Process.*, vol. 29, pp. 2714–2727, Nov. 2019.
[93] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3D pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2500–2509.
[94] K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah, "Social cues in group formation and local interactions for collective activity analysis," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, Feb. 2013, pp. 539–548.
[95] P. A. Tresadern and I. D. Reid, "Camera calibration from human motion," *Image Vis. Comput.*, vol. 26, no. 6, pp. 851–862, Jun. 2008.
[96] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2008, pp. 691–704.
[97] J. Vandoni, E. Aldea, and S. Le Hégarat-Mascle, "Evidential query-by-committee active learning for pedestrian detection in high-density crowds," *Int. J. Approx. Reasoning*, vol. 104, pp. 166–184, Jan. 2019.
[98] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "Detecting conversational groups in images and sequences: A robust game-theoretic approach," *Comput. Vis. Image Understand.*, vol. 143, pp. 11–24, Feb. 2016.
[99] S. Vascon and M. Pelillo, "Detecting conversational groups in images using clustering games," in *Multimodal Behavior Analysis in the Wild*. New York, NY, USA: Academic, 2019, pp. 247–267.
[100] J. Vester, "Estimating the height of an unknown object in a 2D image," M.S. thesis, KTH CSIC, Stockholm, Sweden, 2012.
[101] W. P. Voon, N. Mustapha, L. S. Affendey, and F. Khalid, "Collective interaction filtering approach for detection of group in diverse crowded scenes," *KSII Trans. Internet Inf. Syst.*, vol. 13, no. 2, pp. 912–928, 2019.
[102] R. Wagner, D. Crispell, P. Feeney, and J. Mundy, "4-D scene alignment in surveillance video," 2019, *arXiv:1906.01675*. [Online]. Available: https://arxiv.org/abs/1906.01675
[103] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3D human poses from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2361–2368.
[104] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, 2015, pp. 1299–1302.
[105] W. Wang, W. Lin, Y. Chen, J. Wu, J. Wang, and B. Sheng, "Finding coherent motions and semantic regions in crowd scenes: A diffusion and clustering approach," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014, pp. 756–771.
[106] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.
[107] J. Watson, M. Firman, A. Monszpart, and G. J. Brostow, "Footprints and free space from a single color image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11–20.
[108] *Depression and Other Common Mental Disorders*, WHO Document Prod. Services, World Health Org., Geneva, Switzerland, 2017.
[109] H. Wildenauer and A. Hanbury, "Robust camera self-calibration from monocular images of Manhattan worlds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2831–2838.
[110] B. Xu and G. Qiu, "Crowd density estimation based on rich features and random projection forest," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.
[111] Yang, Gonzalez-Banos, and Guibas, "Counting people in crowds with a real-time network of simple image sensors," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, p. 122.
[112] B. Yuan, A. Wu, and W.-S. Zheng, "Does a body image tell age?" in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2142–2147.
[113] L. Zhang, H. Lu, X. Hu, and R. Koch, "Vanishing point estimation and line classification in a manhattan world with a unifying camera model," *Int. J. Comput. Vis.*, vol. 117, no. 2, pp. 111–130, Apr. 2016.
[114] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 398–407.



MARCO CRISTANI (Member, IEEE) is currently an Associate Professor (Professore Associato) with the Computer Science Department, University of Verona, an Associate Member with the National Research Council (CNR), and an External Collaborator with the Italian Institute of Technology (IIT). He is also the Managing Director of the Computer Science Park, a technology transfer center at the University of Verona. He is or has been the Principal Investigator of several national and international projects, including PRIN and H2020 projects. His main research interests include statistical pattern recognition and computer vision, mainly in deep learning and generative modeling, with application to social signal processing and fashion modeling. On these topics, he has published more than 170 articles, including two edited volumes, six book chapters, 40 journal articles, and 129 conference papers. He has organized 11 international workshops, co-founded a spin-off company, Humatics, dealing with visual analytics. He is a member of IAPR and ELLIS. He is a member of the Editorial Board of *Pattern Recognition* and *Pattern Recognition Letters* journals.

ALESSIO DEL BUE (Member, IEEE) is currently a Tenured Senior Researcher leading the Pattern Analyisis and computer VISion (PAVIS) Research Line, Italian Institute of Technology (IIT), Genoa, Italy. He is a coauthor of more than 100 scientific publications in refereed journals and international conferences. His current research interests include 3D scene understanding from multi-modal input (images, depth, and audio) to support the development of assistive artificial intelligence systems. He is a member of ELLIS. He is a member of the technical committees of important computer vision conferences (CVPR, ICCV, ECCV, and BMVC). He serves as an Associate Editor for *Pattern Recognition* and *Computer Vision and Image Understanding* journals.

VITTORIO MURINO (Senior Member, IEEE) received the Laurea degree in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Genova, Italy, in 1989 and 1993, respectively. He joined the Ireland Research Centre, Huawei Technologies (Ireland) Company Ltd., Dublin, as a Senior Video Intelligence Expert. From 1995 to 1998, he was an Assistant Professor with the Department of Mathematics and Computer Science, University of Udine, Italy. Since 1998, he has been working with the University of Verona, Italy, where he is currently a Full Professor. He was the Chairman of the Department of Computer Science, from 2001, year of foundation, to 2007. From 2009 to 2019, he was the Director of the Pattern Analysis and Computer Vision (PAVIS) Department, Istituto Italiano di Tecnologia, Genova, Italy. He is a coauthor of more than 400 articles published in refereed journals and international conferences. His main research interests include computer vision, pattern recognition, and machine learning, more specifically, statistical and probabilistic techniques for image and video processing for (human) behavior analysis and related applications, such as video surveillance and biomedical imaging. He is a Fellow of IAPR. He is a member of the technical committees of important conferences (CVPR, ICCV, ECCV, ICPR, and ICIP), and a Guest Co-Editor of special issues in relevant scientific journals. He is also a member of the Editorial Board of *Computer Vision and Image Understanding* and *Machine Vision and Applications* journals.

FRANCESCO SETTI (Member, IEEE) is currently an Assistant Professor with the Department of Computer Science, University of Verona, working on the EU-H2020 Project SARAS, and an Associate Researcher with the Institute of Cognitive Science and Technology, Italian National Research Council (ISTC-CNR). He is a coauthor of more than 40 scientific publications in refereed journals and international conferences. His research interests include application of machine learning and artificial intelligence techniques for industrial applications, with attention to the emerging fields of collaborative robotics and reinforcement learning for situation awareness decision making. He serves as an Associate Editor for *Neurocomputing* and *Cognitive Processing*.

ALESSANDRO VINCIARELLI (Member, IEEE) is currently with the University of Glasgow, where he is also a Full Professor with the School of Computing Science and an Associate Academic with the Institute of Neuroscience and Psychology (http://vinciarelli.net). His main research interest includes social signal processing, the domain aimed at modeling analysis and synthesis of nonverbal behavior in social interactions. He has published more than 150 works, including one authored book, and more than 40 journal articles. He is the Co-Founder of Klewel, a knowledge management company recognized with national and international awards. He has been the General Chair of the IEEE International Conference on Social Computing, in 2012, and the ACM International Conference on Multimodal Interaction, in 2017. He is or has been a Principal Investigator of several national and international projects, including the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, a European Network of Excellence (the SSPNet), and more than ten projects funded by the Swiss National Science Foundation and the U.K. Engineering and Physical Sciences Research Council.