

SALSA: A Novel Dataset for Multimodal Group Behavior Analysis

Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, *Member, IEEE*, Ligia Batrinca, Elisa Ricci, *Member, IEEE*, Bruno Lepri, Oswald Lanz, *Member, IEEE*, Nicu Sebe, *Senior Member, IEEE*

Abstract—Studying free-standing conversational groups (FCGs) in unstructured social settings (e.g., *cocktail party*) is gratifying due to the wealth of information available at the **group** (mining social networks) and **individual** (recognizing native behavioral and personality traits) levels. However, analyzing social scenes involving FCGs is also highly challenging due to the difficulty in extracting behavioral cues such as target locations, their speaking activity and head/body pose due to crowdedness and presence of extreme occlusions. To this end, we propose **SALSA**, a novel dataset facilitating multimodal and Synergetic sociAL Scene Analysis, and make two main contributions to research on automated social interaction analysis: (1) SALSA records social interactions among 18 participants in a natural, indoor environment for over 60 minutes, under the *poster presentation* and *cocktail party* contexts presenting difficulties in the form of low-resolution images, lighting variations, numerous occlusions, reverberations and interfering sound sources; (2) To alleviate these problems we facilitate multimodal analysis by recording the social interplay using four *static surveillance cameras* and *sociometric badges* worn by each participant, comprising the *microphone*, *accelerometer*, *bluetooth* and *infrared* sensors. In addition to raw data, we also provide *annotations* concerning individuals' personality as well as their *position*, *head*, *body orientation* and *F-formation* information over the entire event duration. Through extensive experiments with state-of-the-art approaches, we show (a) the limitations of current methods and (b) how the recorded multiple cues synergetically aid automatic analysis of social interactions. SALSA is available at <http://tev.fbk.eu/salsa>.

Keywords—Multimodal group behavior analysis, Free-standing conversational groups, Multimodal social data sets, Tracking, Speaker recognition, Head and body pose, F-formations, Personality traits.

1 INTRODUCTION

HUMANS are social animals by nature, and the importance of **social interactions** for our physical and mental well-being has been widely acknowledged. Therefore, it is unsurprising that social interactions have been studied extensively by social psychologists in a



Fig. 1. Analysis of round-table meetings (left, adopted from *Mission Survival* dataset [6]) has been attempted extensively—orderly spatial arrangement and sufficient separation between persons enable reliable extraction of behavioral cues for each person from such scenes. Exemplar SALSA frame (right) showing FCGs—varying illumination, low resolution of faces, extreme occlusions and crowdedness makes AASI highly challenging (best-viewed in color and under zoom).

variety of contexts. Fundamental research on social interactions was pioneered by Goffman [1], whose *symbolic interaction perspective* explains society via the everyday behavior of people and their interactions. Face-to-face conversations are the most common form of social interactions, and **free-standing conversational groups** (FCGs) denote small groups of two or more co-existing persons engaged in ad-hoc interactions [2]. FCGs emerge naturally in diverse social occasions, and interacting persons are characterized by mutual scene locations and poses resulting in spatial patterns known as F-formations [3]¹. Also, social behavioral cues like how much individuals speak and attend to others are known to be correlated with individual and group-specific traits such as Extraversion [4] and Dominance [5].

Automated analysis of social interactions (AASI) has been attempted for over a decade now. The ability to recognize social interactions and infer social cues is critical for a number of applications such as surveillance, robotics and social signal processing. Studying unstructured social scenes (e.g., a cocktail party) is extremely challenging since it involves inferring (i) locations and head/body orientations of targets (persons) in the scene, (ii) semantic and prosodic auditory content corresponding to each target, (iii) F-formations and persons interact-

1. A F-formation is a set of possible configurations in space that people may assume while participating in a social interaction.

- Xavier Alameda-Pineda, Ligia Batrinca and Nicu Sebe are with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. (Email: xavier.alamedapineda@unitn.it, ligia.batrinca@unitn.it, niculae.sebe@unitn.it).
- Elisa Ricci, Bruno Lepri and Oswald Lanz are with Fondazione Bruno Kessler, Trento, Italy. (Email: eliricci@fbk.eu, lepri@fbk.eu, lanz@fbk.eu).
- Jacopo-Staiano is with Sorbonne Universités, UPMC, CNRS, LIP6, France (Email: jacopo.staiano@lip6.fr).
- Ramanathan Subramanian is with the Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore. (Email: subramanian.r@adsc.com.sg).

ing at a particular time (addresser and addressee recognition), (iv) attributes such as the big-five personality traits [7] from the above social behavioral cues.

Some of the aforementioned problems have been effectively addressed in controlled environments such as round-table meetings (Fig. 1 (left)), where behavioral cues concerning orderly arranged participants can be reliably acquired through the use of web-cameras and close-talk microphones. However, *all* of them remain unsolved in unstructured social settings, where only distant surveillance cameras can be used to capture FCGs [3], and microphones may be insufficient to clearly recognize the speaker(s) at a given time instant due to scene crowding. We are expressly interested in analyzing FCGs (Fig. 1 (right)) in this work.

To address the challenges involved in analyzing FCGs, we present **SALSA**, a novel dataset² facilitating multimodal and Synergetic social Scene Analysis. In contrast to social datasets studying round-table meetings [6], [8], or examining FCGs on a small scale [9], [10], SALSA contains uninterrupted recordings of an indoor social event involving 18 subjects over 60 minutes, thereby serving as a rich and extensive repository for the behavioral analysis and social signal processing communities. In addition to the raw multimodal data, SALSA also contains position, pose and F-formation annotations over the entire event duration for evaluation purposes, as well as information regarding participants' personality traits.

The social event captured in SALSA comprised two parts: (1) a *poster presentation* session, and (2) a *cocktail party* scene where participants freely interacted with each other. There were no constraints imposed in terms of how the subjects were to behave (*i.e.*, no scripting) or the spatial layout of the scene during the recordings. Furthermore, the indoor environment where the event was recorded was prone to varying illumination and reverberation conditions. The geometry of F-formations is also influenced by the physical space where the social interaction is taking place, and the poster session was intended to simulate a semi-structured social setting facilitating speaker labeling as described in Section 5.4. While it is reasonable to expect observers to stand in a semi-circular fashion around the poster presenter, none of the participants were instructed on where to stand or what to attend. Finally, as seen in Fig. 1 (right), the crowded nature of the scene and resulting F-formations gave rise to extreme occlusions, which along with the low-resolution of faces captured by surveillance cameras, make visual analysis extremely challenging. Also, crowdedness poses difficulties to audio-based analysis for solving problems such as speaker recognition. Overall, SALSA represents the most challenging dataset for studying FCGs to our knowledge.

To alleviate the difficulties in traditional audio-visual analysis due to the challenging nature of the scene, in ad-

dition to four wall-mounted cameras acquiring the scene, *sociometric badges* [11] were also worn by participants to record various aspects of their behavior. These badges include a microphone, an accelerometer, bluetooth and infrared (IR) sensors. The microphone records auditory content that can be used to perform speaker recognition under noisy conditions, while the accelerometer captures person motion. Bluetooth and IR transmitters and receivers provide information regarding interacting persons, and are useful for inferring body pose under occlusions. Cumulatively, these sensors can synergistically enhance estimation of target locations and their head and body orientations, face-to-face interactions, F-formations, and thereby provide a rich description of FCGs' behavior. Through the SALSA dataset, we make two main contributions to AASI research:

- Firstly, we believe that the challenging nature of SALSA will enable researchers to appreciate the limitations of current AASI approaches, and spur focused and intensive research in this regard.
- We go beyond audio-visual analysis for studying FCGs, and show how multimodal analysis can alleviate difficulties posed by occlusions and crowdedness to more precisely estimate various behavioral cues and personality traits therefrom.

The paper is organized as follows. Section 2 reviews prior unimodal and multimodal approaches to AASI, while Section 3 highlights the limitations of traditional AASI approaches and datasets, motivating the need for SALSA. Section 4 describes the SALSA dataset, and the synergistic benefit of employing audio, visual and sensor-based cues is demonstrated via experiments in Section 5. We finally conclude in Section 6.

2 LITERATURE REVIEW

This section reviews the state-of-the-art in social behavior analysis with specific emphasis on methods analyzing FCGs. We will first discuss unimodal approaches (*i.e.*, vision-, audio- and wearable sensor-based) and then describe expressly multimodal approaches.

2.1 Vision-based approaches

Challenges pertaining to surveillance scenes involving FCGs addressed by vision-based methods include detection and tracking of targets in the scene, estimation of social attention direction and detection of F-formations. We describe works that have examined each of the above problems as follows.

Given the cluttered nature of social scenes involving FCGs, detecting and tracking the locations of multiple targets is in itself a highly challenging task. Multi-target tracking from monocular images is achieved through the use of a dynamic Bayesian network in [12]. As extreme occlusions are common in social scenes involving FCGs (see Fig. 1), employing information from multi-camera views can enable robust tracking. The color-based particle filter tracker proposed in [13] for multi-target tracking

2. The SALSA dataset (raw data, annotations and associated code) is available at <http://tev.fbk.eu/salsa>.

with explicit occlusion handling is notable in this regard. Of late, tracking-by-detection [14] combined with data association using global appearance and motion optimization [15] has benefited multi-target tracking. Some works have further focused on identity-preserving tracking over long sequences [16]. These methods assume a sufficiently large number of high-confidence detections which might not always be available with FCGs due to persistent long-term occlusions, although recent works have focused on the detection problem under partial occlusion [17], [18]. Multi-target tracking is further shown to improve by incorporating aspects of human social behavior [19], [20], as well as with activity analysis [21]. However, F-formations are special groups, characterized by static arrangements of interacting persons constrained by their locations and head/body orientations.

Social attention determination in round-table meetings, where high-resolution face images are available, has been studied extensively. Vision-based approaches typically employ head pose as a proxy to determine social attention direction [22]. In comparison, head pose estimation (HPE) from blurry surveillance videos is much more difficult. Visual attention direction of moving targets is estimated by Smith *et al.* [12] using position and head pose cues, but social scenes or occlusions are not addressed here. Some works have exploited the relationship between walking direction and head/body pose to achieve HPE from surveillance videos under limiting conditions where no labeled training examples are available, or under occlusions [23], [14], [24]. Focus-of-attention estimation in social scenes is explicitly addressed by Bazzani *et al.* [25], who model a person’s visual field in a 3D scene using a subjective view frustum.

Recently, the computer vision community has showed some interest in the detection and analysis of dyadic interactions [26], [27] and more general groups [28], [29], [30]. Also, the interest on detecting social interactions and F-formations from video has intensively grown. Cristani *et al.* [31], [32], [33] employ positional and head orientation estimates to detect F-formations based on a Hough voting strategy. Bazzani *et al.* [25] gather information concerning F-formations in a social scene using the Inter-relation pattern matrix. F-formations are modeled as maximal cliques in edge-weighted graphs in [34], and each target is presumed to be oriented towards the closest neighbor but head/body orientation is not explicitly computed. Gan *et al.* [35] detect F-formations upon inferring location and orientation of subjects using depth sensors and cameras. Setti *et al.* [9] propose a graph-cuts based framework for F-formation detection based on the position and head orientation estimates of targets. Based on experiments performed on four datasets (IDIAP Poster [34], Cocktail Party [10], Coffee Break [31] and GDet [25]), their method outperforms six state-of-the-art methods in the literature.

2.2 Audio-based approaches

Studying interactional behavior in unstructured social settings solely using audio or speech-based cues is extremely challenging, as FCGs are not only characterized by speaking activity, but also by non-verbal cues such as head and body pose, gestural and proxemic cues. Furthermore, classical problems in audio analysis become extremely challenging and remain unexplored in crowded indoor environments involving a large number of targets. Indeed, current methodologies for speaker diarization [36], sound source separation [37] or localization [38] address scenarios with few persons. Nevertheless, a few studies on audio-based detection of FCGs have been published. Wyatt *et al.* [39] tackle the problem of detecting FCGs upon recognizing the speaker using temporal and spectral audio features. Targets are then clustered to determine co-located groups on the basis of speaker labels. More recently, FCG detection and network inference is achieved in [40] employing a conversation sensing system to perform speaker recognition, and F-formations are detected based on proximity information obtained using bluetooth sensors.

2.3 Wearable-sensor based approaches

Wearable sensors can provide complementary behavioral cues in situations where visual and speech data are unreliable due to occlusions and crowdedness. Hung *et al.* [41] detect FCGs in social gatherings by measuring motion via an accelerometer. With the increased usage of smartphones, mobile sensor data have also become a viable choice for analysis of social interactions or more complex social systems [42]. Via mobile phones, proximity can be inferred from Wifi and bluetooth [43]. However, the spatial resolution of these sensors is limited to only a few meters, and the co-location of mobile devices does not necessarily indicate a social interaction between the corresponding individuals. Therefore, Catuto *et al.* [44] propose a framework that balances scalability and resolution through a sensing tier consisting of cheap and unobtrusive active RFID devices, which are capable of sensing face-to-face interactions as well as spatial proximity over different scale lengths down to one meter or less. Nevertheless, we note that many of these works address relatively less crowded scenarios, not comparable in complexity to SALSA.

2.4 Multimodal approaches

Multimodal approaches to social interaction have essentially examined small-group interactions such as round-table meetings, and mainly involve audio-visual analysis as detailed below. Examples of databases containing audio-visual recordings and associated annotations are the Canal9 [45], AMI [8], Mission Survival [6], Free Talk [46] and the Idiap WOLF [47] corpora. In addition, the IMADE [48] and UEM [49] technological frameworks for recording of multimodal data to describe social scenes involving FCGs are available. Robotics is one among

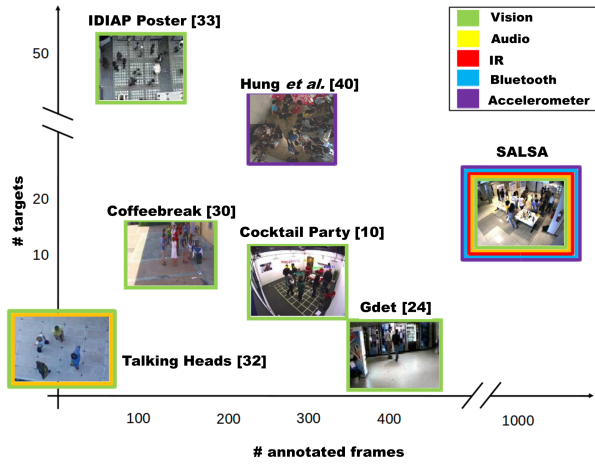


Fig. 2. Existing datasets facilitating F-formation detection. Frame border colors encode sensing modalities.

other applications for which datasets have been published, see [50]. All these data collection efforts have inspired inter-disciplinary research in the field of human behavior understanding and led to the emergence of the social signal processing community [51]. Several of these databases have been utilized for isolating traits relating to an individual (e.g., big-five personality traits) or a group (such as *dominance*). Choudhury and Pentland [11] initiated behavior analysis using wearable sensors by developing the *Sociometer*. Recently, Olguín *et al.* [52] proposed the *Sociometric badge*, which stores (i) motion using an accelerometer, (ii) speech features (rather than raw audio), (iii) position information and (iv) proximity to other individuals using a bluetooth sensor, and (v) face-to-face interactions via an infrared sensor. Sociometric badges have been used to capture face-to-face communication patterns, examine relationships among individuals and model collective and organizational behavior [53], detect personality traits and states, and predict outcomes such as productivity and job satisfaction [54]. Another notable AASI work employing multi-sensory information is that of Matic *et al.* [55], who estimate body orientation and inter-personal distance via mobile data and speech activity to detect social interactions.

3 SPOTTING THE RESEARCH GAP

While human behavior has been studied extensively in controlled settings such as round-table meetings, achieving the same with FCGs is way more difficult as close audio-visual examination of targets is precluded by the crowded and occluded nature of the scene. We carried out an extensive analysis of previous AASI data sets focusing on FCGs; they have mainly been used to address two research problems: (1) Detecting F-formations and (2) Studying individual and group behavior from multiple sensing modalities.

The vast majority of works addressing F-formation detection are vision-based. Fig. 2 presents snapshots of datasets used for F-formation detection, and positions

Dataset	Vision	Audio	IR/Rfid	BT	Accel.
AMI [8]	★	★			
Canal9 [44]	★	★			
Mission Survival [6]	★	★			
Free Talk [45]	★	★			
Idiap Wolf Corpus [46]	★	★			
SALSA	★	★	★	★	★
Sociometric Corpus [52]		★	★		★
Boston Sociometric C. [53]		★	★	★	★
Trento Sociometric C. [54]		★	★	★	★
Matic <i>et al.</i> [55]		★		★	

Fig. 3. Datasets for social interaction analysis: the first five consist on round-table meetings and span over hours, while the last four study social networks/behavior and span over days/months.

them with respect to the number (denoted using #) of annotated frames and scene targets. Among them, SALSA is unique due to its (i) multimodal nature, (ii) extensive annotations available over a long duration and (iii) challenging nature of the captured scene.

Figure 3 depicts the datasets used for individual and group behavioral analysis. While the first group (light-gray) consists of audio-visual recordings spanning over hours acquired under controlled settings, the second group (dark-gray) comprises datasets acquired over days/months for studying social networks and group relationships. SALSA again stands out as it records information from both static cameras and wearable sensors, leading to a previously non-existent and highly informative combination of sensing modalities. This section details some of the limitations of current AASI approaches regarding F-formation detection and individual and group behavior analysis, thereby throwing light on how SALSA can spur critical research in these respects.

3.1 Human tracking and pose estimation

Human tracking and pose estimation in a social context is challenging for several reasons. Firstly, a person’s visual appearance can change considerably across the scene due to camera perspective and uneven illumination, as well as with pose and posture. Secondly, large and persistent occlusions are frequent in such scenes, which corrupt subsequent observations. Thirdly, integrating auditory information to aid localization and orientation estimation is also difficult due to its intermittent nature and the adverse impact of reverberations and interfering sources. Beyond inherent complexity, the state-of-the-art is further challenged when raw audio-visual data cannot be recorded for processing or scene-specific optimization, e.g., due to privacy concerns.

Accurate FCG behavior analysis requires the correct assignment of observations to sources (targets) over the long run. Identity switches during tracking can corrupt the extraction of aggregated features that develop over time to infer personality traits, functional roles, or interaction networks, and long-term identity-preserving multi-target tracking is still unachievable. Furthermore, existing appearance-based pose estimation methods are not adapted to highly cluttered scenes with large and persistent occlusions. Finally, from the computational perspective, multi-target approaches are not able to robustly and efficiently scale to large groups.

3.2 Speech processing

While numerous research studies have attempted speech, speaker and prosodics recognition under controlled conditions, several issues arise when auditory information is captured via mobile microphones in crowded indoor scenes. Firstly, regular indoor environments are prone to reverberations, which adversely affect many sound processing techniques. Secondly, intermittence of the speech signal necessitates speaker diarization prior to processing. Thirdly, the speech signal is also spatially sparse, and source separation techniques are usually required to segment speaker activity. Currently, there are no algorithms addressing source separation or diarization in the presence of a large number of sound sources and uncontrolled conditions. While multimodal approaches have addressed these problems via audio-visual processing, they still cannot work with large groups of people and crowded indoor environments involving unconstrained and evolving interactions.

3.3 F-formation detection

Detecting F-formations in unconstrained environments is a complex task. As F-formations are characterized by mutually located and oriented persons, robust tracking and pose estimation algorithms are necessary. However, both multi-target tracking and head/body pose estimation in crowded scenes are difficult as discussed earlier. Even under ideal conditions, F-formation shapes are influenced by (i) the environment’s layout, *i.e.*, room shape, furniture and other physical obstacles, (ii) scene crowdedness and (iii) attention hotspots such a poster, painting, *etc.* While existing methodologies typically assume that F-formation members are placed on an ellipse, robust F-formation detection requires accounting for the above factors as well. Also, most algorithms are visually driven, and few multimodal approaches exist to this end.

3.4 Inferring personality traits

Works seeking to recognize personality traits from interactive behavior have traditionally relied on the visual and auditory modalities. Target position and pose, prosodic and intonation inference, face-to-face interaction detection, bodily gestures and facial expressions are commonly used for assessing the big-five personality traits. Most prior works study these behaviors in the

context of round-table meetings, where participants are regularly arranged in space, and therefore their positions and body orientations are known *a priori*. Under these conditions, behavior analysis algorithms deliver precise estimates concerning interactional behavior, thereby facilitating personality trait recognition. Nevertheless, there are very few works that studied personality inference from unstructured interactions involving FCGs.

Assessing the personality traits of a large number of interacting persons in crowded scenarios, where the group structure evolves progressively, is a highly challenging task. In such cases, people constantly leave and join groups, and therefore groups are created, split, and merged. An in-depth analysis should take the group dynamics into account in addition to the evolving physical arrangements and occasional conversations. Given that personality inference is a sophisticated and subtle task, the right combination of cues extracted from different modalities can lead to a robust assessment.

3.5 The *raison d’être* of SALSA

Upon analyzing the state-of-the-art in behavior analysis and personality inference, we conclude that: (i) Even if multimodal analysis has been found to outperform unimodal approaches and provide a richer representation of social interplays, some key tasks are not yet addressed in a multimodal fashion, *e.g.* pose estimation and F-formation detection; (ii) Social interactions have been studied under controlled settings, and there is a paucity of methods able to cope with unconstrained environments involving large groups, crowded spaces and highly dynamic interactions, and (iii) Most existing approaches have independently studied the different behavioral tasks – while it is known that bidirectional links between the tasks exist, these links have been rarely exploited. For instance, knowing the head and body orientations of individuals can help in the estimation of F-formations and vice-versa. Similarly, F-formation detection clearly benefits from accurate tracking algorithms, which at their turn can be influenced by the robust detection of F-formations. In order to foster the study of the aforementioned challenges, we recorded SALSA, whose description is presented in the next section.

4 THE SALSA DATA SET

In order to provide a new and challenging evaluation framework for novel methodologies addressing the aforementioned challenges, we introduce the SALSA (Synergetic social Scene Analysis) dataset. SALSA represents an excellent test-bed for multimodal human behavior understanding due to the following reasons. Firstly, all behavioral data were collected in a regular indoor space with the participants only requiring to wear portable and compact sociometric badges which ensured naturalistic social behavior. Secondly, due to the unconstrained nature of the scene, the recordings contain numerous artifacts such as varying illumination,

TABLE 1
Description of the sensors used in SALSA. STFT denotes short-time Fourier transform.

Sensor	Output	Freq. (Hz)
Vision	4 synchronized images	15
Audio	Amplitude stats & STFT	2 & 30
Infra-red	Detected badge's ID	1
Bluetooth	Detected badge's ID	1/60
Accelerometer	Body motion	20

visual occlusions, reverberations or interfering sound sources. Thirdly, the recorded event involved 18 persons: such large social groups have rarely been studied in the behavior analysis literature. These participants did not receive any special instructions or scripts prior to the recording, and the resulting social interactions were therefore free-willed and hedonistic in nature. Finally, the social interplay was recorded via four wall-mounted surveillance cameras and the *Sociometric badges*³ worn by the targets. These badges recorded different aspects of the targets' social behavior such as audio or motion as detailed later. This combination of static cameras and wearable sensors is scarce in the literature, and provides a wealth of behavioral information as shown in Section 5. These four salient characteristics place SALSA in a unique position among the various datasets available for studying social behavior, see Figures 2 and 3.

4.1 Scenario and roles

SALSA was recorded in a regular indoor space and the captured social event involved 18 participants and consisted of two parts of roughly equal duration. The first part consisted of a *poster presentation* session, where four research studies were presented by graduate students. A fifth person chaired the poster session. In the second half, all participants were allowed to freely interact over food and beverages during a *cocktail party*.

It needs to be noted here that while some participants had specific roles to play during the poster presentation session, none were given any instructions on how to act in the form of a script. Consequently, the interaction dynamics correspond to those of a natural social interplay. Obviously, participants with different roles (chair, poster presenter, attendee) are expected to have different interaction dynamics, and these roles were designed to help behavioral researchers working on role recognition.

4.2 Sensors

The SALSA data were captured by a camera network and wearable badges worn by targets. The camera network comprised four synchronized static RGB cameras (1024×768 resolution) operating at 15 frames per second (fps). Each participant wore a sociometric badge during the recordings which is a $9 \times 6 \times 0.5$ cm box equipped with four sensors, namely, a microphone, an infrared (IR) beam and detector, a Bluetooth detector and an accelerometer. The badges are battery-powered and store



Fig. 4. Five annotated F-formations represented via connections between feet positions (crosses) of interacting targets. Corresponding O-spaces are denoted by the colored convex shapes.

recorded data on a USB card without the need for any wired connection, thus enabling natural social interplay. Table 1 presents an overview of the five sensors used.

4.3 Ground truth data

Annotations

In order to fulfill the requirements expected of a systematic evaluation framework, SALSA provides ground-truth annotations, which were performed either manually or semi-automatically over the entire event duration. The annotations were produced in two steps. In the first step, using a dedicated multi-view scene annotation tool, the *position*, *head* and *body* orientation of each target was annotated every 45 frames (3 s). To speed up the annotation process, the total number of targets was divided among three annotators. A target's position, head and body orientation were annotated by a first annotator and then double-checked by the second. Discrepancies between their judgments were resolved by a third annotator. All annotators were clearly instructed on how to perform the annotations. To facilitate the annotation task, markings from the previous annotated frame were displayed so that only small modifications were needed.

In the second step, annotated positions and head/body orientations were used for deducing F-formations. Annotations were again performed every 45 frames and we employed the following criteria for detecting F-formations: an F-formation is characterized by the mutual locations and head, body orientations of interacting targets, and is defined by the convex *O-space* they encompass such that each target has unhindered access to its center. A valid F-formation was assumed if the constituent targets were in one of the established patterns, or had direct and unconstrained access to the O-space center in case of large groups (refer to [31] for details). Figure 4 illustrates five annotated F-formations around four posters (target feet positions are marked with crosses) and corresponding O-spaces. Considering the two groups in the foreground, the F-formation in front of the poster on the right does not include the FCG with two targets on the left, since neither of them have access to the center of the larger group.

3. <http://www.sociometricsolutions.com/>



Fig. 5. Distributions of the big-five personality traits.
Personality data

SALSA also contains big-five personality trait scores of participants to facilitate behavioral studies. Prior to data collection, all participants filled the *Big Five* personality questionnaire [7]. The Big Five questionnaire owes its name to the five traits it assumes as constitutive of personality: *Extraversion*– being sociable, assertive, playful vs. aloof, reserved, shy; *Agreeableness*– being friendly and cooperative vs. antagonistic and fault-finding; *Conscientiousness*– being self-disciplined, organized vs. inefficient, careless; *Emotional Stability*– being calm and equanimous vs. insecure and anxious; and *Creativity*– being intellectual, insightful vs. shallow, unimaginative. In the questionnaire, each trait is investigated via ten items assessed on a 1–7 Likert scale. The final trait scores were computed according to the procedure detailed in [56], and the distributions of these traits over the 18 targets are presented in Figure 5.

5 EXPERIMENTS ON SALSA

This section is devoted to evaluate the performance of state-of-the-art methodologies for different behavioral tasks on SALSA. In order to ensure reproducibility of results, we will provide (i) the dataset, (ii) the software used to produce the obtained results and (iii) a detailed description of the experiments. While the data and the software will be made available online, a complete description of the conducted experiments can be found in the supplementary material.

5.1 Multimodal synchronization

One critical issue when recording with several independent devices is synchronization, as the same event is labeled by independent sensors with different timestamps. In the case of SALSA, we had to synchronize the eighteen badges worn by targets to the camera network. Assuming that there is no drift, we need to find the time-stamp mapping between each sociometric badge and the camera network. This problem reduces to determining the time-shift for each badge so that events are simultaneously observed by the badge and the camera network.

Using the position and body pose annotations, we determined the set of potential infra-red detections, time-stamped with respect to the cameras. By computing the

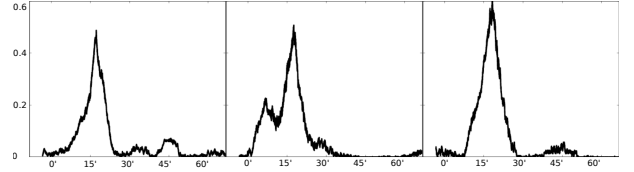


Fig. 6. Synchronization procedure: Similarity scores for badge/target IDs 5, 10 and 16 as a function of the time-shift– a clear peak can be observed for all badges.

similarity score between the potential and actual infra-red detections, we robustly estimated the temporal shift between each badge and the camera network. Computed scores for three of the badges, as a function of the shift are shown in Figure 6. We can observe a clear peak in the badge’s similarity score at the optimal time-shift. Computational details and obtained plots for all badges can be found in the supplementary material.

5.2 Visual tracking of multiple targets

Despite many advances in computer vision research, tracking individuals is still a unsolved problem. In the particular case of SALSA, person tracking is challenging due to the presence of extreme and persistent occlusions. Some targets are difficult to distinguish from others using appearance features, and identity-preserving tracking required for multimodal behavior interpretation is further hindered by non-uniform scene illumination even when multiple views are available. State-of-the-art tracking-by-detection methods feature global appearance optimization [16], but require a sufficiently dense number of high-confidence detections across the whole sequence. However, target detection by itself is extremely challenging in such scenes even if leveraged through learning a set of detectors adapted to different occlusion levels [17]. We therefore considered a sequential Bayesian tracking approach without appearance model adaptation. The Hybrid Joint-Separable particle filter (HJS-PF) tracker [13] was specifically developed for systematic occlusion handling at frame-level, and has been applied to tracking in social scenes [10], [33].

HJS-PF represents targets’ states in the scene based on ground locations, and applies a multi-target color likelihood with a first-order dynamical model to propagate a particle-set approximation of the posterior marginals for each target. In particular, the tracker exploits camera calibration information and a coarse 3D shape model for each target to explicitly model occlusion in the joint likelihood. While exact joint tracking is intractable with increasing number of targets (exponential blow-up induced by the curse-of-dimension), it is shown in [13] that single-target marginals can be updated under explicit occlusion reasoning with quadratic complexity, making the tracking of all 18 SALSA targets feasible. Furthermore, to prevent marginals from overlapping when targets have similar appearance– a frequent failure mode leading to identity switches – a Markov Random Field (MRF) defined over the targets’ positions is added in the

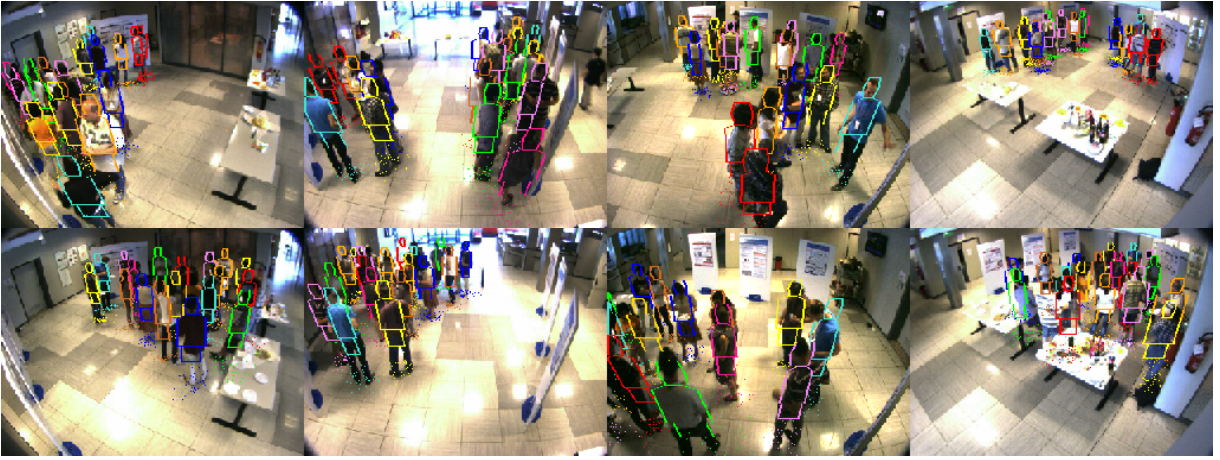


Fig. 7. Tracking on Part 1 - *Poster* (top row) and Part 2 - *Party* (bottom row). Best viewed in color.

propagation. At each HJS-PF iteration, the propagation is solved via message-passing and the update combines HJS-PF likelihoods from each view independently. With a final resampling, the *a posteriori* particle representation is obtained for each target. Details can be found in the supplementary material.

We report HJS-PF tracking results on SALSA following the Visual Object Tracking Challenge (VOT 2013-14) evaluation protocol. The color model for each target was manually extracted from the initial part of the sequence where the target was free of occlusions, prior to tracking. These models were used for the whole sequence and were not re-initialized or adapted during tracking. HJS-PF was initialized for each target with the first available annotation, and tracking was performed at full frame rate (15 Hz) with 320×240 resolution, while evaluation was done every 3 s (or every 45 frames) consistent with the annotations. If the position estimate was over 70 cm from its reference, it was counted as a failure and the tracking of that target was re-initialized at the reference. Otherwise, the distance from the reference was accumulated to compute precision. In Table 2, the average precision (average distance from the references), per-target failure rate (% of failures over 20K annotations), and frames-to-failure count (number of subsequent frames successfully tracked) are reported for the (i) first 30K frames (*Poster*), (ii) the remaining 25K frames (*Party*) and (iii) the total 60 minute recording. Our multi-thread implementation used in these experiments tracks the 18 targets using 50 particles per target at 7 fps on a 3 GHz PC. While overall precision is high considering space dimensionality, low image-resolution and high occlusion rate (cf. last row of table; to our best knowledge no comparable dataset exists for tracking evaluation), a sensible increase in failure rate is observed for the *Party* session. Indeed, in FCGs, persons tend to occupy every available space and exhibit a relaxed body posture such that they are hardly visible in some of the camera views. Also, targets more often bend their bodies to grab food

TABLE 2
Mean tracking statistics and per-target occlusion rates for the four views.

	<i>Poster</i>	<i>Party</i>	All
Precision (cm)	15.2 ± 0.1	20.1 ± 0.1	17.3 ± 0.1
Failure rate (%)	2.6 ± 0.1	9.6 ± 0.3	5.7 ± 0.2
Frames-to-failure	1644 ± 63	439 ± 12	759 ± 21
Occlusion (%)	28,35,22,26	25,28,49,27	27,32,34,27

and beverages, and illumination varies considerably over the scene impeding color-based tracking. However, low failure rate in the *Poster* session where targets arrange themselves in a more orderly manner around posters indicates that occlusion handling is effective with the HJS-PF filter. A snapshot of the tracking results during the *Poster* and the *Party* scenarios is found in Figure 7. Based on these results, we identify some key elements requisite for FCG tracking: (i) perform ground tracking with explicit occlusion handling at frame level, (ii) extract discriminative signatures for each target to resolve identity switches (as in re-identification research), and (iii) learn the illumination pattern of the scene to adapt signatures locally to lighting conditions. These may be cast into a global optimization framework [16] and extended to multi-modal tracking.

5.3 Head and body pose estimation from visual data

The estimation of the head and body pose is still an important research topic in the computer vision community. Specifically, when focusing on estimating the positions and head and body orientation of individuals in FCGs monitored by distant surveillance cameras, several challenges arise due to low resolution, clutter and occlusions. To demonstrate these challenges on SALSA, we considered the recent work of Chen *et al.* [14], which is one of the few methods that jointly compute head and body orientation from low resolution images. In a nutshell, this algorithm consists of two phases. First, Histograms of Oriented Gradients (HoG) are computed from head and body bounding boxes obtained from training data. Then,

TABLE 3
Head and body pose estimation error (degree).

% training data		view 1	view 2	view 3	view 4
10%	Head	45.7 \pm 0.6	47.2 \pm 0.3	48.4 \pm 0.8	49.5 \pm 1.2
	Body	49.3 \pm 0.5	51.6 \pm 0.9	51.2 \pm 0.4	54.6 \pm 0.8
5%	Head	43.6 \pm 0.5	46.2 \pm 0.3	46.4 \pm 0.8	47.5 \pm 0.9
	Body	47.3 \pm 0.5	49.4 \pm 0.5	49.9 \pm 0.5	52.5 \pm 0.7
1%	Head	42.2 \pm 0.4	45.3 \pm 0.3	43.4 \pm 0.8	44.9 \pm 1.5
	Body	45.4 \pm 0.5	47.5 \pm 0.8	48.7 \pm 0.7	51.7 \pm 0.5

a convex optimization problem that jointly learns two classifiers for head and body pose respectively is solved. Importantly, the classifiers are learned simultaneously, imposing consistency on the computed head and body classes so as to reflect human anatomic constraints (*i.e.*, the body orientation naturally limits the range of possible head directions). The approach in [14] leverages information from both annotated and unsupervised data via a manifold term which imposes smoothness on the learned classification functions, typical of semi-supervised learning methods. In our experiments, only labeled data were used for training.

The method proposed in [14] is monocular and considers 8 classes (corresponding to an angular resolution of 45°) for both head and body classification. Therefore, to test it on SALSA, we also considered each camera view separately. In this series of experiments, the target head and body bounding boxes were obtained by manual annotation, and a subset of about 7.5K samples was employed (bounding boxes were not available for targets going out of the field of view, see supplementary material). To compute visual features for both head and body, we used the HoG descriptors. In our tests, a small percentage of the frames (1%, 5%, 10%) were used for training, while the rest were used for testing. For performance evaluation, we used the mean angular error (in degrees) defined in [14] for computing head and body pose estimation accuracy.

Experiments were repeated ten times with random training sets, and corresponding average error and standard deviation are reported in Table 3. Despite many occlusions and the presence of clutter, a state-of-the-art pose classification approach achieves satisfactory performance (maximum error of around one class width). However, it is worth noticing that our experiments were performed with homogeneous training and test data, in contrast with the heterogeneous data employed in [14]. We expect a significant decrease in performance when heterogeneous training data are used for pose estimation. Also, errors observed for head pose are considerably smaller than for body pose over all four camera views—this is because body pose classifiers are impeded by severe occlusions in crowded scenes. Precisely for this reason, previous works on F-formation detection from FCGs [31], [33], [57] have primarily employed head orientation, even though body pose has been widely acknowledged as the more reliable cue for determining interacting persons. We believe that devising a multi-



Fig. 8. Mean speaker recognition accuracy with different methods on SALSA.

view [58] and multimodal [59] approach also employing IR and bluetooth-based sensors for body pose estimation would be advantageous as compared to a purely visual analysis, which was one of the primary motives for compiling the SALSA dataset.

5.4 Speaker recognition

Speaker recognition is a critical and fundamental task in behavior analysis from FCGs. Processing the auditory data in SALSA is challenging for several reasons. First, the recordings were carried out in a regular indoor space prone to reverberations and ambient noise. Second, 18 persons participated in the event and freely interacted with one another and therefore, the audio recordings consist of mixtures of speech signals emanating from different speakers. Third, the sociometric badges only retained part of the time-frequency representation, and thus high-performance speaker recognition is very challenging on this data. Finally, as relative positions of the speakers were constantly evolving, the speaker-to-microphone filter is not only unknown but highly time-varying, and thus very hard to estimate in practice. Indeed, current methods for speaker localization [38], [60] or source separation [37], [61] are not designed for such a complex scenario.

Classical speaker recognition approaches build on Mel Frequency Cepstral Coefficients (MFCCs). Computed from the short-time frequency transform (STFT), MFCC have been shown to achieve a good balance between descriptive power, complexity and dimensionality [62]. Four classifiers, namely, support vector machines with linear (SVM-L) and radial-basis function kernel (SVM-RBF), Gaussian mixture models (GMM) and random forests (RF) were employed for MFCC-based speaker recognition. To create ground-truth labels, we annotated by visual inspection the ID of the speaker within a group of five persons interacting over a 15-minute duration during the poster session. This yields one 6-class (5 persons and silence) classification problem. We chose this group as the camera perspective and resolution allowed for reliable vision-based annotation. In addition to the straightforward strategy of feeding the badge-specific MFCCs to classifiers, we also concatenated the MFCCs extracted from all badges at every frame to deal with time varying speaker-microphone relative locations. In

TABLE 4
F-formation detection with ground-truth data.

	Prec.	Head Rec.	F1	Prec.	Body Rec.	F1
HVFF lin [31]	0.56	0.72	0.63	0.59	0.74	0.67
HVFF ent [32]	0.63	0.77	0.69	0.66	0.8	0.73
HVFF ms [33]	0.58	0.73	0.64	0.61	0.76	0.68
GT [57]	0.90	0.75	0.82	0.89	0.70	0.78
GC [9]	0.80	0.85	0.82	0.82	0.85	0.83

this way, we use badge’s information when the speaker is active and when it is not active. We refer to these two strategies as “No Fusion” (NF) and “Frame-based Fusion” (FBF) respectively.

Mean speaker recognition accuracies obtained upon five-fold cross validation are presented in Figure 8. We observe that the FBF strategy systematically outperforms the NF strategy. Focusing on FBF, we observe that RBF-SVM and random forests perform similarly, while doing much better than GMM and outperforming linear SVM. Also, as GMMs are known to be less effective in higher dimensional spaces, we computed principal components so as to 90% variance in the FBF setting.

In light of these results, we outline directions for future work. First, due to the crowded nature of scenes involving FCGs, auditory analysis is highly challenging. Due to ambient noise, reverberations and multiple sound sources, recognizing the speaker from the badge audio data is challenging *per se*. Nevertheless, performance increase observed with the FBF strategy suggests that a multi-modal approach can be effective, where tracking and pose estimates can facilitate multi-microphone based speech analysis. Finally, examining the badge data closely, most non-zero STFT coefficients are in the first nine frequency bins ($< 300\text{Hz}$). Therefore, algorithms attempting speaker recognition on SALSA should design features to exploit this frequency range.

5.5 F-formation detection

Detecting F-formations by visual observing crowded scenes is a challenging task. Several factors such as low video resolution, occlusions and complexities of human interactions hinder robust and accurate F-formation detection. We first considered four state-of-the-art vision-based approaches for individuating FCGs in SALSA. We adopted (i) Hough voting [31] (HVFF-lin), (ii) its non-linear variant [32] and (iii) multi-scale extensions [33] (denoted as HVFF-ent and HVFF-ms), (iv) the game-theoretic (GT) approach [57] and (v) the graph cut (GC) approach [9] as associated codes are publicly available⁴.

These approaches take the targets’ positions and head pose as input, and compute F-formations independently for each frame. In particular, the Hough-voting methods work by generating a set of virtual samples around each target. These samples are candidate locations for the O-space center. By quantizing the space of all possible

locations, aggregating samples in the same cell and finding the local maxima in the discrete accumulation space, the O-space centers and F-formations therefrom are identified. Oppositely, in the graph-cut algorithm, an optimization problem is solved to compute the O-space center coordinates.

We first evaluated the above F-formation detection approaches using ground-truth position and head and body pose annotations, and considered all the annotated frames. F-formation estimation accuracy is evaluated using precision, recall and F1-score as in [31]. In each frame, we consider a group as correctly estimated if at least $T \cdot |G|$ of the members are correctly found, and if no more than $1 - T \cdot |G|$ non-members are wrongly identified, where $|G|$ is the cardinality of the F-formation G and $T = 2/3$. Results are reported in Table 4. Even the most accurate approach, *i.e.* the graph-cut method, only achieves a F1-score of about 0.83, clearly demonstrating the need of devising more sophisticated algorithms for detecting F-formations in challenging datasets such as SALSA. Moreover, it is worth noting that our results are consistent with the observations in previous works such as [31], *i.e.*, using the body pose is more advantageous than using head orientation for detecting a group of interacting persons.

In a second series of experiments, we evaluated the graph-cut approach using *automatically* estimated head and body orientations from the multi-sensor badge data. Specifically, we considered annotations for the target positions, and estimated head and body pose from visual data with the method proposed in [14]. In these experiments, HoG features extracted from head and body crops for the four camera views were concatenated and provided as input to the classifiers. In this series of experiments, we only considered a subset of frames where *all* the targets were in the camera field of view. To train the coupled head-body pose classifier, we used 1% of the available samples as training data. Experiments were repeated ten times and the average performance is reported. We also integrated information from IR and audio sensors. Audio and IR are sparse observations, whilst visual data are available at every time-stamp. The likelihood of target n addressing m was estimated from audio and IR data. The maximum likelihood points to the person with whom m (the addressee) is more likely to interact. The audio and IR observations correspond to the direction of the addresser (n). For integrating multiple angles, we simply considered their weighted average. Weights were tuned so as to maximize algorithm performance on a small validation set.

The results of our experiments are reported in Fig. 9. Clearly, when the head and body pose are automatically computed from visual analysis, the performance significantly decreases with respect to the use of ground-

4. <http://profs.sci.univr.it/~cristanm/ssp/>
<http://www.iit.it/it/datasets-and-code/code/gtgc.html>

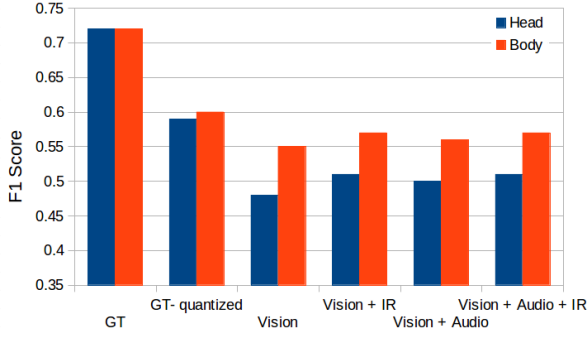


Fig. 9. F-formation detection results (F1 score). Head and body pose were automatically estimated from visual, infra-red and audio data.

truth (GT) annotations⁵. Furthermore, by combining information from multiple modalities, we obtain a modest improvement with respect to using only visual data. Specifically, while jointly employing visual and infra-red data is advantageous with respect to exclusively employing the visual data, the integration of audio sensors provides minimum benefit. Finally, it is worth noting that a decrease in performance with respect to the ground-truth is also due to the angle quantization process. This can be observed by comparing the four leftmost bars in Fig. 9. Therefore, casting head and body pose estimation as a classification task (as typical of related works, where four/eight pose classifiers are used) appears to be insufficient for robustly detecting F-formations. Instead of considering classifiers, our experiments suggest that a better strategy entails casting head and body pose estimation as a regression task.

5.6 Interaction networks and personality traits

The SALSA behavioral data also allow for investigating the relationship between social dynamics and higher-level behavioral determinants such as personality traits. In this section, different from traditional works that have correlated audio-visual behavioral cues with the big-five traits, we show how the sociometric badge data can also be utilized for the same. To this end, we built three networks based on i) infra-red (IR) hits; ii) audio correlations⁶; and iii) group compositions from video-based F-formation annotations (GT). To account for the dynamics, we selected windows of 60 and 120 s, and proceeded to build a multimodal graph for each window. For infra-red, the graph representing the n -th window has an edge between the target pairs whose badges detected an infra-red hit during the time period defined by n . For audio, we employed the correlations and added an edge between two targets if the corresponding correlation value

5. Note that the accuracies with GT data reported in Fig. 9 are different from those presented in Table 4 since only a subset of frames is used in these experiments. In these considered frames, the scene is crowded since all 18 targets are inside the field of view.

6. We computed the correlation between the badge STFT coefficients (normalized by the energy of the recording badge).

is above a threshold (empirically set to 0.95). From the video data, we added an edge between two targets when they were detected as being part of the same group. Furthermore, we also built a static multimodal graph which encoded the entire sequence (equivalent to setting the window’s duration equal to the sequence’s duration). From these networks, we extracted three basic classes of structural characteristics, i.e., *centrality*, *efficiency*, and *transitivity*, and investigated how these characteristics are related to personality traits.

Inspired by previous studies [63], [64], we extracted the three standard measures of centrality proposed by Freeman: *degree*, *betweenness*, and *closeness* centrality [65]. These centrality measures can be divided into two classes: those based on the idea that the centrality of a node in a network is related to how close the node is to the other nodes (e.g., *degree* and *closeness* centrality), and those based on the idea that central nodes stand between others playing the role of intermediary (e.g., *betweenness* centrality). Furthermore, we computed the network constraint [66] for each individual; this measure provides an indication on how much the target’s connections are connected with one another.

We also computed *nodal* and *local* efficiency for each node in the networks. The concept of *efficiency* [67] can be used to characterize how close to a ‘small world’ a given ego-network is. Small world networks are a particular kind of networks that are highly clustered, like regular lattices, and have short characteristic path lengths like random graphs [68]. The use of efficiency is justified by the hypothesis that the rate at which information flows within the network is influenced to some degree by the personality of the ego.

Finally, we extracted the *transitivity* measure, which provides an indication of the clustering properties of the graph under analysis. Based on triads, i.e., triples of nodes in which either two (*open*) or three (*closed*) nodes are connected by an edge, *transitivity* is defined as the ratio of the number of closed triads to the number of graph triads. In [63], Extraversion was found to negatively correlate with *local transitivity*, while McCarty and Green [69] found that agreeable and conscientious persons tend to have well-connected networks.

We conducted a preliminary statistical analysis (summarized in Table 5) on the features derived from the interaction graphs described above, and investigated their associations with the personality data provided by SALSA. We only report associations that are significant at $p < .05$, unless otherwise stated. Unfortunately, we did not find statistical significant correlations between the personality traits and the auditory features. This issue will be subject of further investigation.

Extraversion, a personality trait lying in the tendency to behave in a way to engage and attract other people, and hence usually activated in situations such as social gatherings, was found to be significantly associated ($R = 0.53$) with the standard deviation of the degree centrality computed on the 60 s dynamic infra-red network. In

TABLE 5

Significant Pearson correlations between the big-five traits and IR and GT network features (* denotes $p < .01$).

Trait	Feature	R
Extr.	Betw. Median Dyn-120-IR	.53
	Degree Std Dyn-60-IR	.53
Agre.	Degree Median Dyn-120-IR	.48
	Nodal Eff. Median Dyn-120-IR	.49
Cons.	Local Eff. Stat-IR	.52
	Nodal Eff. Median Dyn-60-GT	.49
	Nodal Eff. Median Dyn-120-GT	.56
	Trans. Mean Dyn-60-IR	.54
Em. St.	Trans. Median Dyn-60-IR	-.53
Crea.	Betw. Stat-GT	.49
	Clos. Mean Dyn-60-IR	-.49
	Degree Mean Dyn-60-GT	-.47
	Degree Mean Dyn-120-GT	-.50
	Loc. Eff. Stat-IR	-.51
	Nod. Eff. Stat-GT	-.49
	Trans. Stat-GT	-.73*

other words, more extraverted targets appear to establish face-to-face interactions of variable duration, and thus engage with groups of diverse cardinality within the 1-minute windows under analysis. The expansion of the time window to 120 s provided additional insights: the higher a target scored on the Extraversion trait, the higher the median of his/her *betweenness* centrality ($R = 0.54$). This suggests that the extravert targets in SALSA tended to act as *brokers*, i.e., they served as connectors between clusters of people who had fewer face-to-face interactions.

Also, more agreeable subjects in SALSA were found to have a tendency towards engaging in face-to-face interactions with a higher number of people within highly connected clusters. Agreeableness was found to be significantly associated with the median degree centrality ($R = 0.48$) and the median nodal efficiency ($R = 0.49$) as computed on the 120 s graphs. Emotional Stability was found to be negatively associated with median transitivity ($R = -0.53$) on the 60 s infra-red graphs—this indicated that neurotic persons in SALSA tended to engage face-to-face during unbalanced interaction events.

Regarding Conscientiousness, several significant associations were found. In particular, from the 60 s dynamic interaction networks built on infra-red data, we noted that conscientious subjects tended to participate in densely connected groups, as indicated by the positive association with mean transitivity ($R = 0.54$). Thus, within the groups that naturally formed in the SALSA context, conscientious subjects took part mainly in those groups where the participants engaged more with each other. This fact is further confirmed by the significant association found on the static graph built on infra-red data between this trait and local efficiency ($R = 0.52$). Interestingly, the median nodal efficiency extracted from the dynamic graphs built upon group annotations consistently shows similar associations with this trait ($R = 0.56$ using a 120 s window, $R = 0.49$ using 60 s).

The social psychology literature does not offer many

quantitative studies regarding the Creativity (or Openness to Experience) trait. Given this lack of information, the associations detected were remarkable, and deserve deeper investigation. This trait seemed to be associated with: i) local efficiency ($R = -0.51$) computed on the static face-to-face interaction network; ii) mean closeness centrality ($R = -.49$) computed on the 60 s graphs built on infra-red; iii) betweenness centrality ($R = 0.49$), nodal efficiency ($R = -0.49$), and transitivity ($R = -0.73$, $p < .01$) computed on the static group interaction network; and iv) mean degree in both 60 s ($R = -0.47$) and 120 s ($R = -.5$) graphs built from group annotations. Hence, creative persons seemed to participate overall in smaller and less connected networks than conservative targets.

While the above analyses present correlations between some of the behavioral cues that can be extracted from the SALSA data and high-level personality traits, we believe the entire gamut of information available can enable the study of both individual and group-level traits (e.g., *dominance*). Also, unlike round-table meetings, which typically have an agenda based on which participants assume certain roles that may not relate to their actual personality, SALSA captures hedonistic and free-wheeling social interactions, which even if challenging to analyze, can provide a wealth of information about one's native behavior and personality.

6 CONCLUSIONS AND FUTURE WORK

Via extensive experiments, we have demonstrated how SALSA represents a rich but challenging dataset for analysis of FCGs. Vision-based analysis for target tracking, head and body pose estimation and F-formation detection evidenced the shortcomings of state-of-the-art methodologies when posed with cluttered scenes with persisting and extreme occlusions. However, additional sensors available as part of the sociometric badge were found to be helpful in cases where visual analysis was difficult—in particular, (i) the infra-red sensor which indicates both the proximity and body pose of the interacting counterpart was found to improve F-formation detection, (ii) both IR and visual cues were found to have significant correlations with the big-five personality traits, and (iii) improved speaker recognition with multi-badge speech data indicates the promise of additionally utilizing visual and accelerometer data to this end.

Future research directions include: (a) developing new methodologies for robust audio processing in cluttered environments with many dynamic targets, (b) utilizing the bluetooth and accelerometer data for F-formation detection and personality trait recognition, and (c) designing tracking and head/body pose estimation algorithms capable of exploiting multimodal data. Given the extensive raw data and accompanying annotations available for analysis and benchmarking, we believe SALSA can spur systematic and intensive research to address the highlighted problems in a multimodal fashion in the near future. Evidently, SALSA would serve as a precious resource for the computer vision, audio processing, social

robotics, social signal processing and affective computing communities among others.

7 ACKNOWLEDGEMENTS

This work was supported by the EC project ACANTO, by the cluster project Active Ageing at Home and by the HCCS grant from A*STAR Singapore.

REFERENCES

- [1] E. Goffman, *Encounters: Two studies in the sociology of interaction*. Bobbs-Merrill, 1961.
- [2] —, *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Glencoe, Illinois: Free Press, 1963.
- [3] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge: CUP Archive, 1990, vol. 7.
- [4] M. Ashton, K. Lee, and S. Paunonen, "What is the central feature of extraversion? Social attention versus reward sensitivity," *J. Person. Social Psych.*, vol. 83, no. 1, pp. 245–52, 2002.
- [5] E. Rosa and A. Mazur, "Incipient status in small groups," *Social Forces*, vol. 58, no. 1, pp. 18–37, 1979.
- [6] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion - a systematic study," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 443–455, 2012.
- [7] O. John and S. Srivastava, "The Big Five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research*, 1999, pp. 102–138.
- [8] I. McCowan et al., "The AMI meeting corpus," in *Int. Conf. on Meth. and Tech. in Beh. Res.*, 2005.
- [9] 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, 2015.
- [10] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: Towards socially and personality aware visual surveillance," in *ACM Work. on Multi. Perv. Vid. Anal.*, 2010.
- [11] T. Choudhury and A. Pentland, "Sensing and modeling human networks using the sociometer," in *IEEE Int. Symp. Wear. Comp.*, 2003.
- [12] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *TPAMI*, vol. 30, no. 7, pp. 1212–1229, 2008.
- [13] O. Lanz, "Approximate bayesian multibody tracking," *TPAMI*, vol. 28, no. 9, pp. 1436–1449, 2006.
- [14] C. Chen and J.-M. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *CVPR*, 2012.
- [15] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *TPAMI*, vol. 33, no. 9, pp. 1806–1819, 2011.
- [16] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Multi-commodity network flow for tracking multiple people," *TPAMI*, vol. 36, no. 8, pp. 1614–1627, 2014.
- [17] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, "Handling occlusions with Franken-classifiers," in *ICCV*, 2013.
- [18] C. Wojek, S. Walk, S. Roth, and B. Schiele, "Monocular 3d scene understanding with explicit occlusion reasoning," in *CVPR*, 2011.
- [19] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *ICCV*, 2009.
- [20] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *TPAMI*, vol. 34, no. 5, pp. 1003–1016, 2012.
- [21] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," in *NIPS*, 2010.
- [22] R. Stiefelhagen, "Tracking focus of attention in meetings," in *ICMI*, 2002.
- [23] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *ICCV*, 2011.
- [24] A. K. Rajagopal, R. Subramanian, E. Ricci, R. L. Vieri, O. Lanz, N. Sebe, et al., "Exploring transfer learning approaches for head pose classification from multi-view surveillance images," *IJCV*, vol. 109, no. 1–2, pp. 146–167, 2014.
- [25] 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 Systems*, vol. 30, no. 2, pp. 115–127, 2013.
- [26] M. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari, "Detecting people looking at each other in videos," *IJCV*, vol. 106, no. 3, pp. 282–296, 2014.
- [27] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *TPAMI*, vol. 34, no. 12, pp. 2441–2453, 2012.
- [28] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Discovering groups of people in images," in *ECCV*, 2014.
- [29] M. Eichner and V. Ferrari, "We are family: Joint pose estimation of multiple persons," in *ECCV*, 2010.
- [30] Y. Yang, S. Baker, A. Kannan, and D. Ramanan, "Recognizing proxemics in personal photos," in *CVPR*, 2012.
- [31] 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 *BMVC*, 2011.
- [32] F. Setti, H. Hung, and M. Cristani, "Group detection in still images by F-formation modeling: A comparative study," in *WIAMIS*, 2013.
- [33] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-scale F-formation discovery for group detection," in *ICIP*, 2013.
- [34] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in *ICMI*, 2011.
- [35] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli, "Temporal encoded F-formation system for social interaction detection," in *ACM Multimedia*, 2013.
- [36] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *TASLP*, vol. 20, no. 2, pp. 356–370, 2012.
- [37] R. Badeau and M. Plumbley, "Multichannel HR-NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain," in *WASPAA*, 2013.
- [38] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *TASLP*, vol. 22, no. 6, pp. 1082–1095, 2014.
- [39] D. Wyatt, T. Choudhury, J. Bilmes, and J. A. Kitts, "Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science," *ACM Trans. Intel. Sys. Techn.*, vol. 2, no. 1, 2011.
- [40] C. Luo and M. C. Chan, "Socialweaver: Collaborative inference of human conversation networks using smartphones," in *ACM Conf. Emb. Net. Sen. Sys.*, 2013.
- [41] H. Hung, G. Englebiene, and L. Cabrera-Quiros, "Detecting conversing groups with a single worn accelerometer," in *ICMI*, 2014.
- [42] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Pers. Ubi. Comp.*, 2006.
- [43] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland, "Sensing the 'health state' of a community," *IEEE Perv. Comp.*, vol. 11, no. 4, pp. 36–45, 2012.
- [44] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J. Pinton, and A. Vespignani, "Dynamics of person-to-person interactions from distributed rfid sensor networks," *PLOS ONE*, 2010.
- [45] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *IEEE Int. Conf. Aff. Comp. Intel. Inter.*, 2009.
- [46] E. Kurtic, B. Wells, G. J. Brown, T. Kempton, and A. Aker, "A corpus of spontaneous multi-party conversation in bosnian serbo-croatian and british english," in *Int. Conf. Lang. Res. Eval.*, 2012.
- [47] H. Hung and G. Chittaranjan, "The IDIAP wolf corpus: exploring group behaviour in a competitive role-playing game," in *ACM Multimedia*, 2010.
- [48] Y. Sumi, M. Yano, and T. Nishida, "Analysis environment of

conversational structure with nonverbal multimodal data," in *ICMI*, 2010.

- [49] K. Mase, Y. Sumi, T. Toriyama, M. Tsuchikawa, S. Ito, S. Iwasawa, K. Kogure, and N. Hagita, "Ubiquitous experience media," *IEEE Multimedia*, vol. 13, no. 4, pp. 20–29, 2006.
- [50] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Cech, K. Kulkarni, A. Deleforge, and R. P. Horaud, "Ravel: An annotated corpus for training robots with audiovisual abilities," *JMUI*, vol. 7, no. 1-2, pp. 79–91, 2013.
- [51] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vis. Comp.*, vol. 27, no. 12, pp. 1743–1759, 2009.
- [52] D. Olguín Olguín, B. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Trans. on Sys., Man and Cyber. B*, vol. 39, no. 1, pp. 43–55, 2009.
- [53] B. Lepri, J. Staiano, G. Rigato, K. Kalimeri, A. Finnerty, F. Pianesi, N. Sebe, and A. Pentland, "The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations," in *IEEE Int. Conf. on Soc. Com.*, 2012.
- [54] D. Olguín Olguín and A. Pentland, "Sensor-based organizational design and engineering," *Int. Jour. of Org. Des. and Eng.*, 2010.
- [55] A. Matic, V. Osmani, A. Maxhuni, and O. Mayora, "Multi-modal mobile sensing of social interactions," in *Int. Conf. Perv. Comp. Tech. Heal.*, 2012.
- [56] M. Perugini and L. Di Blas, *The Big Five Marker Scales (BFMS) and the Italian AB5C taxonomy: Analyses from an emic-etic perspective*. Hogrefe & Huber Publishers, 2002.
- [57] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "A game theoretic probabilistic approach for detecting conversational groups," in *ACCV*, 2014.
- [58] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition," *IEEE TIP*, vol. 23, no. 12, pp. 5599–5611, 2014.
- [59] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *ACM MM*, 2015.
- [60] X. Alameda-Pineda and R. Horaud, "Vision-guided robot hearing," *IJRR*, vol. 34, no. 4-5, pp. 437–456, 2015.
- [61] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational em algorithm for the separation of moving sound sources," in *IEEE WASPAA*, 2015.
- [62] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993.
- [63] S. Wehrli, "Personality on social network sites: An application of the five factor model," *ETH Zurich Sociol.*, 2008.
- [64] P. A. Gloor *et al.*, "Towards honest signals of creativity identifying personality characteristics through microscopic social network analysis," *Procedia-Social and Beh. Sci.*, vol. 26, pp. 166–179, 2011.
- [65] L. Freeman, "Centrality in social networks: Conceptual clarification," *Sensor Networks*, 1979.
- [66] R. S. Burt, *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992, vol. 5.
- [67] V. Latora and M. Marchiori, "Economic small-world behavior in weighted networks," *Eur. Phys. Jour. B-Cond. Mat.*, 2003.
- [68] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, 1998.
- [69] C. McCarty and H. Green, "Personality and personal networks," *Sunbelt XXV*, 2005.



Xavier Alameda-Pineda received M.Sc. in mathematics, in telecommunications and in computer science. He worked towards his Ph.D. in mathematics and computer science in the Perception Team at INRIA, and obtained it from Université Joseph Fourier in 2013. He is currently a Post-Doctoral fellow at the University of Trento. His research interests are multimodal machine learning and signal processing for scene analysis.



Jacopo Staiano is a researcher at LIP6, UPMC - Sorbonne Universités. He received his PhD from University of Trento in 2014. He was visiting scholar at the Ambient Intelligence Research Lab, Stanford University, at the Human Dynamics Lab, MIT Media Lab and at Telefonica I+D. His latest research efforts are on mobile computing and the economics of personal data. He received the Best Paper Award at ACM UBICOMP 2014.



Ramanathan Subramanian received a PhD in Electrical and Computer engineering from the National University of Singapore in 2008. He is currently a Research Scientist at the Advanced Digital Sciences Center, Singapore. His research interests span human-centered computing, brain-machine interfaces, human behavior understanding, computer vision, and multimedia processing.



Ligia Batrinca received the PhD degree in Cognitive and Brain Sciences in 2013 from the University of Trento, Italy. During her PhD, she was a visiting student at the Institute for Creative Technologies, USC. She is a Post-Doctoral fellow at the University of Trento. Her research interest are in the area of automatic human behavior analysis, including affective and social behaviors analysis and personality detection.



Elisa Ricci is an assistant professor at University of Perugia and a researcher at Fondazione Bruno Kessler. She received her PhD from the University of Perugia in 2008. She has since been a post-doctoral researcher at Idiap, Martigny and the Fondazione Bruno Kessler, Trento. She was also a visiting student at University of Bristol during her PhD. Her research interests are mainly in the areas of computer vision and machine learning.



Bruno Lepri is a researcher at FBK and affiliate researcher at MIT Media Lab. At FBK, he is leading the Mobile and Social Computing Lab and he is vice-responsible of the Complex Data Analytics research line. In 2009, he received his Ph.D. in Computer Science from the University of Trento. His research interests include multimodal human behavior understanding, automatic personality recognition, network science, and computational social science.



Oswald Lanz received a Masters in Mathematics and a Ph.D. in Computer Science from University of Trento in 2000 and 2005 respectively. He then joined Fondazione Bruno Kessler, and post his tenure in 2008, now heads the Technologies of Vision research unit of FBK. His research interests are mainly in the areas of computer vision and more recently in machine learning for vision. He holds two patents on video tracking.



Nicu Sebe is a professor in the University of Trento, Italy, leading the research in the areas of multimedia information retrieval and human behavior understanding. He was General Co-Chair of the IEEE FG Conference 2008 and ACM Multimedia 2013, and Program Chair of the International Conference on Image and Video Retrieval in 2007 and 2010 and ACM Multimedia 2007 and 2011. He is Program Chair of ECCV 2016 and ICCV 2017. He is a fellow of IAPR.