Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features

Cigdem Beyan, *Member, IEEE,* Francesca Capozzi, Cristina Becchio, and Vittorio Murino, *Senior Member, IEEE*

***Abstract*—The coordination of a leader with group members is very important for an effective leadership given that this figure is the person who actually manages the team members to achieve a desired goal. Investigating the leadership and especially the leadership style is a prominent research topic in social and organizational psychology. However, this is a new problem in social signal processing which can actually make valuable contributions by analyzing multi-modal data in a more effective and efficient way. In this work, we identify the leadership style of an emergent leader (i.e. the leader who naturally arises from a group, not designated) as autocratic or democratic. The proposed method is applied to a dataset in-the-wild, in other words there is no role-playing, which is novel for this problem. Multiple kernel learning (MKL) using multi-modal nonverbal features is utilized to predict leadership styles which proved to achieve better predictions as compared to traditional learning methods. Thanks to MKL and a simple heuristic proposed, the best performing features are also identified, showing that better predictions can be reached only using those features. Additionally, correlation analysis between the extracted nonverbal features and the results of social psychology questionnaire is also performed. This shows that significantly high correlations exist for speaking activity based and prosodic nonverbal features.**

***Index Terms*—Leadership style, emergent leader, social signal processing, nonverbal features, small group interactions, multiple kernel learning, in-the-wild.**

I. Introduction

SOCIAL Signal Processing (SSP) is a relatively recent field which aims to analyze human behaviors i.e social interactions, in an automatic way using some areas of computer science e.g. speech processing, computer vision, machine learning, etc. Social interactions are the fundamental of human life and also the main research track for social psychology. Even though social interaction studies in psychology have a very deep background, the automatic analysis of interactions is still an issue given that the traditional methods in psychology are based on manual processing which is very labor intensive and time consuming [1].

An organization is rich in terms of social interactions. For instance, a meeting environment contains many discussions, problem solving and decision making. In an organization, usually a leader is designated. This leader guides the group activities, influences the group and shows his/her dominance and control over the group members [2]. An emergent leader (EL), on the other respect, is a person who naturally appears as the leader by showing these characteristics during a social interaction such as during a meeting [2]. Detection of ELs is a well-investigated topic in organizational behavioral research although a recent problem in SSP [3]. For an organization, the leader and leader' coordination with group members is very crucial, and an effective leader can direct the group members to achieve a desired success [4]. Thus, identifying the ELship skill of a person in early stages (such as while hiring a person) is useful for an organization [5]. Leadership has been investigated from many perspectives. Much work was also allocated to understand leadership styles [6]. In fact, different leadership styles can affect the performance and productivity of group members which originate the effectiveness and success of an organization [4].

In this work, we concentrate on prediction of two leadership styles: autocratic leadership and democratic leadership of an EL (perceived leader). **Autocratic leadership** is directive, tends to exert strong control over group conversation and decision making process, hence could elicit a centralization of the activities of the group. **Democratic leadership** is participative, tends to value the involvement of group members in the conversation and decision making processes such that it favors more interaction among the group members [7]–[9].

Social interactions are based on verbal and nonverbal (gaze, speaking activity, facial expressions, body activity, etc.) communications. So far, many studies in social psychology (such as [10]) and in SSP (such as [2], [11]–[14]) showed the effectiveness of nonverbal cues for several different tasks in small group interactions. In this study, inspired from ELship and related concepts, automatically extracted nonverbal cues from multiple modalities are used for prediction of leadership styles.

The nonverbal features (NF) are modeled using Multiple Kernel Learning (MKL) which to the best of our knowledge has never been applied for leadership style detection (was applied for EL detection in [12] but only using visual nonverbal cues). Given that we are using different types of features and modalities, MKL can actually perform better than many other machine learning methods as it is able to use different kernels for different feature subsets that have different notion of similarity [15], [16]. Different with [12], we show that MKL can be used like a feature selection algorithm such that the average kernel weights per feature determine the relative classification importance of a feature. By determining the best performing features, the prediction accuracy of even a single

C. Beyan and V. Murino are with Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, 16163, Italy, V. Murino is also with Department of Computer Science, University of Verona, Verona, Italy (e-mail: cigdem.beyan@iit.it; vittorio.murino@iit.it).

F. Capozzi is with Department of Psychology, in McGill University, Montreal, QC, 1205, Canada (e-mail: francesca.capozzi@mcgill.ca).

C. Becchio is with Cognition, Motion and Neuroscience Unit, Istituto Italiano di Tecnologia (IIT), Genova, 16152, Italy and Department of Psychology, University of Turin, Torino, 10124, Italy (e-mail: cristina.becchio@iit.it).

Manuscript received xx xx, 2017; revised xx xx, xx.

kernel classification algorithm (which frequently performs worse than MKL) is improved.

To the best of our knowledge, this work constitutes the most detailed study to date on automatic prediction of leadership styles in small group meetings using multi-modal nonverbal cues. The contributions of this work can be listed as follows.

- For the first time, identification of leadership styles without any role playing (in-the-wild) is investigated.
- A new dataset (publicly available[1]) for leadership style detection is introduced (different with the annotations presented in [17] which were for ELship only).
- A comprehensive survey on leadership style identification and related research are presented.
- For the first time, for prediction of leadership styles, various modalities are utilized individually and all together. This allows to analyze which NFs extracted from these modalities perform better than any other. Furthermore, the correlations between the NFs and the results of questionnaire, which imply leadership styles, are found.
- MKL is applied to predict the leadership styles (for the first time in this work) which generally demonstrates improved results as compared to popular machine learning methods. Additionally, it is showed that MKL can be used as a feature selection method.

The rest of this paper is organized as follows. The previous studies about leadership style detection and a comprehensive review about small group interactions are given in Section II. In Section III, we summarize our approach, the dataset definition including the questionnaire used, and the data annotations are described. In Section IV, the NFs used are described and the methods applied to extract them are presented. The prediction of leadership style using different computational methods are described in Section V. The experimental analysis including leadership style prediction results and the correlation analysis are reported in Section VI. Finally, the paper is concluded with discussions and future work in Section VII.

II. Related Work

In this section, first, the SSP studies which address prediction of leadership styles (LS) are particularly reviewed. The datasets, nonverbal features (NFs) and the learning methods that they utilized are discussed and the main differences between our study are highlighted. Later, various small group interactions were examined in terms of the NFs utilized, the learning methods applied and the correlation analysis performed in SSP and social psychology literatures.

The SSP literature is limited regarding detection of different LS in small group interactions. One of the most related study to ours is [1] which discriminated the individually considerate (defined as the leaders who pay attention to their followers and listen them effectively [18]) and the autocratic leaders. Unlike our study, in [1], only NFs extracted from speaking activity (Speaking-Act) were used. The LS were based on role playing. In detail, the oldest group member was selected as the leader of the group. In our study there is no role playing and leaders naturally exert their dominance and control (i.e. ELs, see [12], [17], [19] for more information). The dataset used in [1] had 34 group discussions such that 16 of them have an autocratic leader. Hence, they had a balanced data (unlike ours) which makes the classification problem less challenging (see [5] for more information). Differently, the prediction of the LS was performed using logistic regression and the training model was learnt using the features of leaders only. Whereas in our study, prediction was performed using leaders' NFs only in addition to the prediction using NFs of all participants. The results [1] showed that, individually considerate leaders started speaking more often while others spoke, used short utterances more often, changed their speech loudness more and spoke less than authoritarian leaders. On the other hand, authoritarian leaders spoke more and had longer turns. Furthermore, the most predictive NFs were detected as: change in single speaking energy, speaking time, short utterances and interruptions.

In [20], posture based NFs (such as left-arm-up, both-arms-up, right-arm-up, etc.) were extracted using wearable motion sensor. Similar to [1], half of the leaders were instructed to perform a considerate leadership while other half behaved as an autocratic leader. The posture mirroring was used as a measure to differentiate LS such that individually considerate leaders showed more posture mirroring than authoritarian leaders while the posture classes: left-arm-up, both-arms-up and right-arm-up were the most frequent classes.

As a following work of [1] and [20], Feese et al. [6] extracted nonverbal cues from body motion to differentiate individually considerate and autocratic leaders (the same dataset presented in [1]). In that study [6], wearable motion sensors were used to extract NFs such as face-touch and arm-crossed. It was shown that, individually considerate leaders mimicked their followers' nodding, face-touch behaviors and posture changes more often as compared to authoritarian leaders. Although, using wearable sensors to quantify mimicry in terms of LS was novel and perhaps performed accurately, using them might cause unnatural body motion.

The group conversational behaviors were investigated for the meetings having autocratic leader, participative leader (egalitarian) or free-rein (a leader who allows group members to make the decision) in [21]. In that study [21], a data corpus which included different roles such as project manager (who was the designated leader (DL)), user interface specialist, marketing expert, industrial designer were utilized. Different with the traditional way which is extracting NFs per participants, the group level features i.e. group patterns were also extracted. Group patterns (presented the group as a whole without taking the identity of the interactions into account) and patterns only belong to DL (could only be applied to the meetings having a DL) were used separately. Latent Dirichlet Allocation (LDA) was used to model these patterns using audio-based NFs only. Although applying an unsupervised algorithm can be an advantage, since their setting needed to know who the leader was (to extract corresponding features), annotation of group participants as a leader or not-a-leader was still necessary. By having different roles assigned to participants and extracting the cues particularly for DL, that work [21] is different than our work which investigates the LS of the perceived leader.

[1] https://www.iit.it/pavis/datasets/leadershipCorpus

In [22], ELship was mentioned as a LS where the leader arises from a group of equal status people. This can be seen complementary with our study such that in each meeting there is a DL and the EL (perceived leader) was annotated by the external observers who were not aware of existence of DL and thus a DL might be labeled as an EL (see Section III for more information). However, it is very important to note that all the analysis performed in our study is regarding to EL.

For identification of LS, the main differences between related works and our study can be listed as follows.

- The existing studies used unnatural scenarios meaning that the leader was either designated [21] or playing a leader role with the specific LS that was assigned [1], [6], [20]. However, we are interested in predicting LS of ELs (perceived leader) who show up naturally due to their dominance, influence and control over other group members.
- Various computational methods have been applied but unlike us, none of the studies applied MKL which is actually useful especially if different modalities and various types of features are used [16].
- The majority of the work used audio NFs and showed that Speaking-Act was an accurate cue to predict LS [1], [21]. On the other hand, visual NFs, particularly, mirroring a certain body posture or body motions were also used [6], [20]. Here, we investigate various audio, video and audio-visual NFs while most of the features were never utilized for LS detection before.
- Differently, we use two types of datasets in our analysis. Although they belong to the same meetings, the way they are processed are different since: *i)* the NFs are extracted independently (such as the modeling of visual focus of attention (VFOA) which requires supervision), *ii)* the annotation of them are also performed differently. Lastly as shown in [23], *iii)* analyzing meeting segments and whole meetings matter because there is never a continuous flow in a meeting and in different moments of a meeting, different leaders with different styles may emerge.
- Unlike other studies, we perform correlation analysis between NFs extracted and the results obtained from social psychology questionnaire which show high correlations for some of the NFs.
- In contrast to studies which performed the analysis using the features extracted from leaders only [1] and studies considered specific (but related) NFs for leaders and different features for non-leaders [21], we perform predictions using three classes (autocratic, democratic and not-a-leader) and two classes (autocratic and democratic) while the same NFs are defined for everybody.

A. NFs & Learning Methods Utilized in SSP

In this section, a comprehensive review of SSP works such that NFs and learning methods investigated during small group interactions is presented. Particularly, leadership style (LS) prediction and the most related topics, i.e. emergent leadership (EL), personal traits (PT; mostly extraversion and locus of control), investigation of meetings in terms of group interactions (G), and dominance (D) are examined. Dominance is analyzed separately since there are many works specifically about it, it is much more related to leadership and LS compared to other personal traits such that autocratic leaders are more dominant than democratic leaders [1]. In Tables I, II, III, and IV, the learning methods used and the audio, visual, audio-visual NFs utilized for different tasks are given, respectively.

TABLE I: Computational methods used

Methods Used: **References**
Logistic regression: [1], [6], [24], [25]
Latent dirichlet allocation: [21], [26], [27]
Rule-based: [2], [11], [13], [19], [23], [28]–[32]
Rank-level fusion: [2], [11], [19], [23], [33], [34]
Collective classification: [2], [23]
Gaussian mixture model: [29]
Hidden markov model: [35]–[37]
"Likelihood ratio" based classifier: [38]
Support vector machine: [2], [12], [13], [17], [23]–[25], [27], [33], [38]–[45]
Naive bayes: [20], [27], [45]
Boosting: [40], [41], [46]
Fisher linear discriminant analysis: [40], [41]
Nearest mean classifier: [40], [41]
Granger causality: [47], [48]
Multiple kernel learning: [12]

This review does not contain the role recognition (except the LS which was defined as a role such as in [1], [21]) studies as we are interested in social interactions in-the-wild and particularly the leadership of the EL (perceived leader).

In overall, various supervised and unsupervised learning methods were used for different tasks. Support Vector Machines (SVM) was one of the most popular approach which was applied for various tasks. Rule-based classifier was mainly applied for ELship. The MKL was presented for the first time in [12] and proved to be the state of the art method for EL detection.

As seen, the most frequently used NF group was Speaking-Act which also showed that audio NFs was preferred more than visual NFs. In terms of task variety Speaking-Act based features, particularly, speaking length, speaking turn, and successful interruptions were utilized more than other features. Although it is hard to generalize, *i)* speaking turn, visual activity (VA) and VA while speaking/listening usually performed better for dominance identification [22], *ii)* audio, audio-visual and visual focus of attention based (VFOA) NFs generally performed better for ELship detection [2], [12], [17], [32], *iii)* VFOA-based, VFOA while speaking based and the Speaking-Act based features were most successful cues for LS prediction (see Section VI), *iv)* Speaking-Act features, energy, pitch, VA-based features and visual dominance ratio were more popular for personal trait differentiation and *v)* group conversational features and group looking features were particularly used for investigation of group interactions when the group was considered as a whole without taking into account the individual interactions.

B. NFs Used in Social Psychology

Psychologists have used nonverbal behavior to analyze the vertical dimension in social interactions which includes

TABLE II: Audio NFs used in the literature for the given tasks

Audio Nonverbal Features: **References**	**Tasks**
1. Speaking activity: [25], [32], [36], [37], [44], [49], [50]	D, EL, G, PT
1.1. Speaking turn (w/wout utterances): [1], [2], [13], [19], [21]–[24], [26], [28], [29], [31]–[35], [43], [45], [46], [51]	D, EL, G, LS, PT
1.2. Speaking length (e.g. total, mean, maximum, in a range, Hellinger length, etc.): [2], [13], [19], [21]–[24], [26], [28]–[36], [38], [40]–[43], [45], [46], [51]	D, EL, G, LS, PT
1.2.1. Single speaking length (e.g. total, maximum, etc.): [1], [30], [52]	D, LS
1.2.2. Multiple speaking length: [1]	LS
1.3. Turn duration statistic (e.g. total, average, minimum, maximum, interquartile range, histogram etc.): [1], [2], [13], [19], [22]–[24], [28], [31]–[33], [38], [42], [43], [46], [51]	D, EL, G, LS, PT
1.4. Overlapped speech (e.g. total, average, fraction, etc.): [2], [22], [25], [44], [46]	D, EL, PT
1.5. Turn taking order: [2], [22], [23], [30]	D, EL
1.6. Successful interruptions: [1], [2], [13], [19], [21]–[23], [26], [28]–[32], [34], [35], [38], [40]–[43], [46], [51]	D, EL, G, LS, PT
1.7. Unsuccessful interruptions: [1], [21], [22], [26], [29], [34]–[36], [51]	D, G, LS
1.8. Being successfully interrupted: [34], [42], [45]	D, G
1.9. Being unsuccessfully interrupted: [34], [45]	D, G
1.10. The successful interruption-to-being interrupted ratio: [42]	D
1.11. Speaker floor grabs: [28], [40]–[43], [46]	D, PT
1.12. Silence (e.g. fraction): [25], [38]	G, PT
1.13. Non-overlapped speech (e.g. fraction): [38]	G
1.14. Being back-channeled: [26], [29], [35]	G
2. Centrality	
2.1. Speaking first after another speaker: [22], [34], [46]	D, EL, PT
3. Prosodic features	
3.1. Energy (e.g. total, spectral flatness, variation, maximum, etc.): [2], [13], [22], [23], [25], [30], [31], [33], [34], [37], [44], [46], [48], [49], [51]	D, EL, G, PT
3.1.1 Single speaking energy (e.g. total, maximum, average, change, etc.): [1], [13], [24], [30]	D, LS, PT
3.2. Pitch (e.g. variation, maximum, mean, etc.): [2], [22]–[25], [37], [44], [49]	D, EL, G, PT
3.3. Rhythm: [22]	D
3.4. Spectral features (e.g. formants, bandwidths, spectrum intensity, etc.): [22], [44]	D, PT
3.5. Speaking rate: [37]	G
4. Group conversational features	
4.1. Group speaking length (e.g. total, as a distribution measure): [21], [22], [26], [29], [35], [38], [45], [51]	D, G, LS
4.2. Group speaking turns (e.g. total, as a distribution measure, skew): [21], [22], [26], [29], [35], [38], [45], [51]	D, G, LS
4.3. Group speaking interruption (e.g. total, as a distribution measure, skew)	
4.3.1 Group successful interruptions (e.g. total, as a distribution measure): [21], [22], [22], [26], [29], [35], [38], [45], [51]	D, G, LS
4.3.2. Group unsuccessful interruptions (e.g. total, as a distribution measure): [21], [22], [26], [29], [35], [45], [51]	D, G, LS
4.3.3 Group successful interruptions-to-turns ratio: [21], [22], [38], [45], [51]	D, G, LS
4.3.4. Group unsuccessful interruptions-to-turns ratio: [21], [22], [45], [51]	D, G, LS
4.4. Group overlapped speech (e.g. fraction): [21], [22], [26], [29], [35], [51]	D, G, LS
4.5. Group silence (e.g. fraction): [21], [22], [26], [29], [35], [45], [51]	D, EL, G, LS
4.6. Group non-overlapped speech (e.g. fraction of it for two, three people): [21], [22], [26], [29], [35], [51]	D, G, LS
4.7. Group egalitarian measures	
4.7.1. Group speaking length egalitarian: [38]	G
4.7.2. Group speaking turns egalitarian measure: [38]	G
4.7.3. Group speaking interruptions egalitarian measure: [38]	G
4.8. Group back-channels (e.g. total, skew): [26], [29], [35]	G

dominance, status, power and leadership [2], [3], [22]. NFs such as total speaking time, interruptions, gazing, smiling, touching, body positions were used to infer social verticality in social interactions [54]. In this section, we particularly focus on the social psychology studies examined NFs for ELship (including dominance as it is a subscale of ELship) and LS.

Regarding audio based NFs, an investigation of ELship using meeting scenarios composed of eight or nine members was presented in [55]. In that study [55], the highest correlation with ELship was found for the relative total time that each participant speaks. In [56], the total speaking time and the average turn duration were found correlated with leadership. Visual dominance (the ratio of the percentage of looking while speaking divided by the percentage of looking while listening) also showed a positive correlation with ELship in [57]. The analysis performed on meetings composed of four-unacquainted participants in [58] showed that the dominant people spoke the most and thus gained more control over the group and the group decision. In [54], ELs showed more gazing, more nodding, and lowered eyebrows, demonstrated less self-touching but more touching to others while their tone of voice was more variable, had faster speech rate, and lower pitch.

Using video based NFs, in [59], arm and shoulder movements were found to be the most important NFs for ELship while gesticulation of arms and shoulders were significantly correlated with gaze, head and facial agreement.

The relationship between LS and social and aggressive dominance were examined in [60]. Their findings presented that, there was a higher correlation between leadership and social-dominance. Social dominant people looked at others more while speaking and used more gestures, received more

TABLE III: Visual NFs used in the literature for the given tasks

Visual Nonverbal Features: **References**	**Tasks**
1. Visual activity (head activity only e.g. using object tracking, body activity only e.g. using object tracking, weighted motion energy image (e.g. entropy, median, etc.), hand activity (e.g. total, standard deviation, histogram, etc.)): [2], [11], [13], [22]–[25], [31], [33]–[37], [43], [44], [47]–[51], [53]	D, EL, G, PT
1.1. Visual activity turns (w/wout short movements): [2], [11], [13], [23], [24], [28], [31], [33], [34], [43], [51]	D, EL, PT
1.2. Visual activity length (e.g. total): [2], [11], [13], [23], [24], [28], [33], [34], [43], [51]	D, EL, PT
1.3. Visual activity turn duration (e.g. total, histogram, average): [2], [11], [13], [23], [24], [28], [31], [33], [43]	D, EL, PT
1.4. Visual activity floor grabs (e.g. total, etc.): [11], [28], [43]	D
1.5. Visual activity unsuccessful interruptions: [11], [35], [36]	D, G
1.6. Visual activity successful interruptions: [13], [22], [28], [33]–[35], [35], [43], [51]	D, EL, G
2. Behavioral mimicry	
2.1. Face touch (e.g. total): [6]	LS
2.2. Arm closed (e.g. total): [6]	LS
2.3. Arm diagonal (e.g. total): [6]	LS
2.4. Hands gesticulating (e.g. total): [6]	LS
2.5. Fidgeting (head, hand, arm, body, etc. (e.g. total)): [6], [44], [49]	LS, PT
2.6. Change in posture (e.g. total): [6], [20]	LS
2.7. Head nodding (e.g. total): [6]	LS
3. Visual focus of attention (gaze): [22]	D, EL
3.1. Visual attention received (e.g. total, minimum, maximum, etc.)	
3.1.1. Visual attention received (looked by at least 1 person w/wout mutual engagement): [12], [17], [32], [34], [39], [51], [52]	D, EL, PT
3.1.2. Visual attention received (looked by at least 2 persons w/wout mutual engagement): [12], [17], [25]	EL, PT
3.2. Visual attention given (w/wout mutual engagement (e.g. total)): [12], [17], [25], [32], [34], [39], [51]	D, EL, PT
3.3. Attention quotient: [12], [17], [32]	EL
3.4. Attention center: [12], [17], [32], [34]	D, EL
3.5. Visual attention turns (e.g. total): [34], [51]	D
3.6. Visual attention given to a non-person: [12], [17]	EL
3.7. Initiating a mutual engagement (e.g. total, standard deviation): [12], [17]	EL
3.8. Intercurrent time between the initiation of mutual engagement (e.g. total, standard deviation): [12], [17]	EL
3.9. Group looking features	
3.9.1. People gaze (e.g. fraction): [26], [35]	G
3.9.2. Convergent gaze (e.g. fraction): [26], [35]	G
3.9.3. Mutual gaze (e.g. fraction): [12], [17], [25], [26], [35]	EL, G, PT
3.9.4. Shared gaze (e.g. fraction): [26], [35]	G
3.9.5. Gaze skew: [26], [35]	G
4. Gestures and facial expressions: [22], [41], [53]	D, EL
5. Group visual activity features (head, body, motion energy image): [25]	PT
5.1. Only one person moves (e.g. total): [25]	PT
5.2. More than two people move (e.g. total): [25]	PT
5.3. Still motion: [25]	PT

frequent and longer-lasting glances from the other participants. In contrast, aggressively dominant people attempted to interrupt more, and looked at others less while listening.

In conclusion, the psychology literature has found that specific NFs were correlated with ELship and LS. These findings and the studies given in Section II-A (which were also motivated by psychological studies) provide the supporting evidence of the NFs used in this study.

III. Overview of the Approach and Dataset

We tackle the leadership style prediction (LS) problem using two types of audio-visual data. One of them includes short segments of meetings (each lasts approx. 5 minutes) such that each includes EL and his/her LS annotations. The other involves the holistic meetings (12-30 minutes) and the corresponding LS annotations are based on social phycology questionnaire. For each person in a meeting (or in a meeting segment), audio-based, video-based and audio-visual NFs are automatically extracted. We use these NFs to predict the different LS using supervised learning. In detail, the training data represented by NFs is modeled using kernel based learning methods. The testing data, which is represented by the same type of NFs, is then classified using the learnt model. The classification result can be democratic leader, autocratic leader or not-a-leader. Additionally, the correlation between NFs and questionnaire results are found. Lastly, we apply feature selection which potentially determines the features that perform the best LS prediction. This proposed approach is illustrated in Figure 1.

The leadership dataset [17] used in this study contains 16 meeting sessions. Whole dataset is 393 minutes while individual meeting sessions last 12 to 30 minutes. Each meeting session is composed of the same gender, unacquainted four-participants (44 females and 20 males with average age of 21.6 with 2.24 standard deviation). For each meeting session, there are five videos. Four videos are from frontal cameras (with resolution of 1280×1024 pixels and frame rate of 20 fps) that captured each participant individually. The other video was recorded using a standard camera (with resolution of 1440×1080 pixels and frame rate of 25 fps) to capture the whole scene which was used for EL annotation in [17]. Audio (sample rate=16 kHz) was recorded with four wireless lapel microphones, each connected to one frontal camera. The plan of the set up can be seen in data acquisition box of Figure 1.

The four participants in a meeting are performing either "winter survival" or "desert survival" [61] tasks which are

TABLE IV: Audio-Visual NFs used in the literature for the given tasks

Audio-Visual Nonverbal Features: References	**Tasks**
1. Visual dominance ratio	
1.1. Ratio between looking at others while speaking and looking at others while another person is speaking: [22], [24], [32], [34], [51], [52]	D, EL, PT
1.2. Ratio between looking at others while speaking and looking at others while not speaking: [22], [24], [27], [32], [34], [52]	D, EL, PT
2. Visual activity while speaking/listening: [40], [50]	D
2.1.Visual activity turns (w/wout short movements): [11], [28], [43]	D
2.2. Visual activity length (e.g. total, average): [11], [28], [43]	D
2.3. Visual activity turn duration (e.g. total): [11], [28], [43]	D
2.4. Visual activity interruptions (e.g. total): [11], [28], [43]	D
2.5. Visual activity floor grabs (e.g. total): [11], [28], [43]	D
3. Visual focus of attention while speaking/not speaking (i.e. listening)	
3.1. Looking someone (e.g. total): [24], [25], [27], [32], [34]	D, EL, PT
3.2. Being looked	
3.2.1. Being looked by one person (looked by at least 1 other persons): [24], [27], [32], [34]	D, EL, PT
3.2.2. Being center of attention (looked by at least 2 other persons): [32]	EL
4. Gesticulation while speaking/listening: [40]	D
5. Using extracted audio (Table II) and visual nonverbal (Table III) features altogether: [2], [13], [23], [24], [26], [27], [31], [35], [35], [37], [44], [48]	D, EL, G, PT

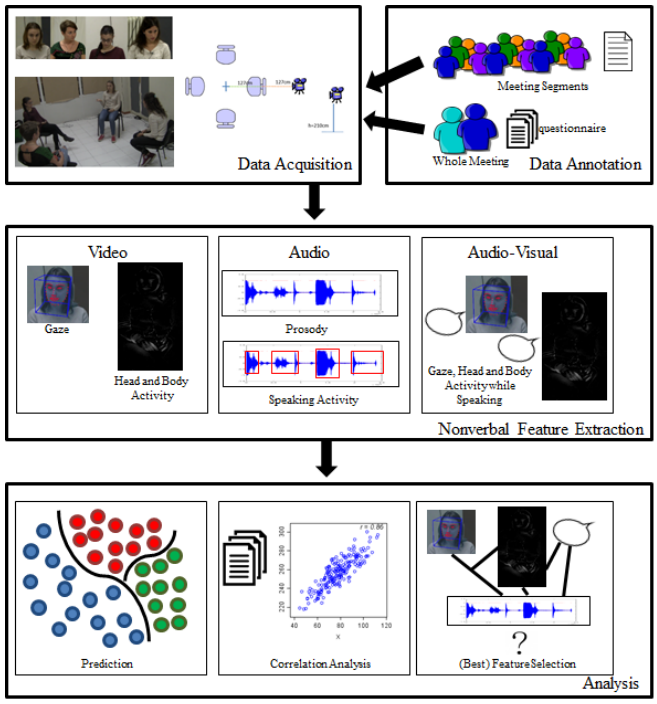

Fig. 1: The overview of the proposed approach.

the most common tasks in small group decision making, dominance and leadership. The details of these survival tasks can be found in [17]. In addition to that, the evaluation of participants' performance in the group task and more details (e.g. how people behave, speak etc.) regarding the meetings are given in the supplementary material.

A. Questionnaire

The SYstematic method for the Multiple Level Observation of Groups (SYMLOG) questionnaire [62], [63] is a tool which considers dominance vs. submissiveness, acceptance vs. non-acceptance of task orientation of established authority, and friendliness vs. unfriendliness. It can be used as a tool for external observation of a group interaction [64]. We use the SYMLOG results (dominance InterClass Correlation (ICC) = 0.866, task-orientation ICC = 0.569, friendliness ICC = 0.722; $p < 0.001$) obtained from two external judges to evaluate the extracted NFs in terms of correlation (Section VI-E) and for annotation of holistic meetings i.e. LS impression per each EL (Section III-C).

For LS impression of holistic meetings, we are based on participants' relational styles which are measured by the friendliness sub-scale of SYMLOG. ELs with higher scores on SYMLOG friendliness are defined as displaying a democratic LS, whereas ELs with lower scores are defined as displaying an autocratic LS. A t-test confirmed that the difference between democratic ELs and autocratic ELs on friendliness scores is reliable ($p=0.04$).

B. Annotation of Meeting Segments

The dataset includes data annotations for 75 small meeting segments which are obtained by dividing 16 meeting sessions into small segments which lasts 5 minutes on average [17]. This dataset can be used mainly for two tasks: *i)* detecting ELs in a meeting environments such as applied in [12], [17] and *ii)* identification of leadership styles which correspond to the detected ELs. The former task is out of scope of this paper and the latter task is investigated for the first time in this paper. The EL and LS annotations were performed by 50 observers in total. Each observer annotated 12 meeting segments on average and each meeting segment was labeled by 8 annotators while no more than one segment which belongs to the same meeting was annotated by the same observer.

Regarding LS, the annotations include three categories: democratic, autocratic and not-a-leader. 66 out of 75 segments have a consensus regarding LS. Therefore, only 66 meeting segments were used. For the autocratic LS annotation, in the 5 out of 20 meeting segments there is a 100% agreement while in the 15 out of 20 meeting segments there is 73% agreement on average. For the democratic LS annotation, in the 5 out of 46 meeting segments there is a 100% agreement while in the 39 out of 46 meeting segments there is 72% agreement. Four out of 16 meeting sessions were also discarded (makes

17 meeting segments in total) for the analysis performed with audio (Section VI-A), audio-visual (Section VI-C) and fusion of audio and video-based NFs (Section VI-D) since they have audio problems. Additionally, the average of the autocratic leaders' distribution and the average of the democratic leaders' distribution in the same meeting are 0.74 and 0.26, respectively with standard deviation equal to 0.27. Out of 16 meetings, in six meetings, all the segments were annotated as democratic LS, in one meeting all the segments were annotated as autocratic LS, and in two meetings the LS annotations were half and half.

C. Annotation of Holistic Meetings

The leadership style of an EL in a given holistic meeting is determined in two steps. First, the EL is found by applying majority voting to the EL annotations of meeting segments (see [17] for more information). It is to be noted that SYMLOG is not used to find the ELs. Then, as mentioned in Section III-A, using the SYMLOG friendliness score, the LS (autocratic or democratic) of the EL, found in the first step, is determined.

The agreement between the LS annotation of meeting segments and LS annotation of holistic meetings is also examined. When majority voting (note that this is not the same with EL annotation which is discussed in the previous paragraph) is applied to the LS annotations of the meeting segments for each meeting, except two meetings out of 16 meetings, it is possible to find a single LS per a meeting. In seven out of 14 meetings the majority voting results are the same with the LS annotations of the corresponding holistic meetings.

As mentioned in Section II, this dataset [17] also includes designated leaders (DLs) but in our analysis we do not take this into account and all analysis performed are based on the annotations of ELs. But it is worth mentioning that a DL may appear as an EL. As given in [17], in the 58 out of 75 meeting segments, the EL annotated by the 50 observers is also the DL. Similarly, in 12 out of 16 holistic meetings, the EL inferred by questionnaire is also the DL. It is not surprising that the majority of the DLs are ELs as well since DLs were defined by a questionnaire (filled by participants two months before the group task, note that this is different from the SYMLOG filled by two external observers given in Section III-A), which evaluates the dominance, which is a measure for leadership. Therefore, DLs were not determined randomly. For meeting segments, we observed that, different leaders can emerge even at the beginning of the meetings. In 4 out of 16 meetings, the EL at the beginning of the task is different than DL. More importantly, the DLs and the existence of DLs were not known by the 50 annotators who decided the ELs and their LS for each meeting segment and also by the two additional external observers who filled the SYMLOG whose results are used to determine the LS for holistic meetings.

IV. Nonverbal Feature (NF) Extraction

In this section, the description of the audio-based, video-based, audio-visual NFs used and the methods utilized to extract them are presented. In total 68 different NFs were extracted. All features (except features based on a ratio) were normalized by the length of the corresponding meeting (or meeting segment) since the lengths of the meetings are variable.

Although the methods used to extract NFs involve slight modifications compared to previous works, the NFs used are standard (given that they have been used for various SSP tasks, see Section II-A). For many times, it has been shown that these features are robust especially for EL detection [2] and dominance [28]. For instance, leaders and dominant participant received more frequent and longer lasting glances from others, looked at others more while speaking, were more active (head and body activity), were more talkative and had longer speaking turns [2], [28]. In this work, it is also presented that many NFs perform well for LS prediction without the necessity to resort to other elaborated features.

A. Audio NFs

The audio NFs include: *i)* speaking activity based features (Speaking-Act) and *ii)* prosodic features (based on energy and pitch, referred as Prosodic). These features are described in Table V with their abbreviations.

1) Speaking Activity based NFs: To extract the Speaking-Act based features, first, speaker diarization was applied. By applying speaker diarization, input audio was segmented into homogenous parts in terms of speakers. In other words, the audio was labeled in terms of "who spoke when" [65]. For speaker diarization, there are many algorithms proposed (such as in [65]) while this is still an active topic in speech processing [22]. We applied the most traditional speaker diarization algorithm which was also applied in [1] for LS detection. This algorithm is based on energy such that a speaking segment of a participant p_i is determined if the energy difference between p_i's energy value and the mean value of the other participants is greater than a threshold. After a binary Speaking-Act vector per participant was obtained (which includes the speaking segments and non-speaking segments of the participant), this vector was de-noised by merging the segments of the same speaker within one seconds. Lastly, the results of de-noised Speaking-Act vectors were checked manually since the algorithm used is prone to error if the optimal thresholds is not set. In total 15 features were extracted using the final Speaking-Act vector (Table V).

2) Prosodic NFs: The prosodic features used includes energy and pitch. These features were computed only when a participant p_i is speaking like applied in [37].

Speaking energy was extracted using the root mean square amplitude of the audio signal over a sliding time window. This window was taken as 40 milliseconds with a 10 milliseconds time shift (as applied in [13], [33], [34]). Speaking pitch was calculated using the algorithm in [66]. This algorithm was chosen because it is robust to acoustic noise even when signal-noise-ratio of the speech is poor such as if the microphone is not close to the speaker. This method uses SIFT algorithm [67] which was applied by many other works such as [37], [68]. From energy and pitch, in total 22 features were extracted (Table V).

TABLE V: Audio NFs for participant p_i.

Speaking-Act based NFs	**Abbreviation**
The total speaking length of p_i when at least one other participant is also speaking.	$TMSL_i$
The total speaking length of p_i when nobody is speaking.	$TSSL_i$
Ratio between $TMSL_i$ and $TSSL_i$.	$FMSSL_i$
The total number of speaking turns of p_i without utterances (a turn lasts minimum 2 seconds).	TST_i
The average speaking turn durations of p_i.	$ASTD_i$
The total number of successful interruptions of p_i: p_i starts talking when p_j is speaking and p_j finishes his/her turn before p_i does.	TSI_i
The total number of unsuccessful interruption of p_i: p_i starts talking when p_j is speaking when p_i finishes his/her turn p_j is still speaking.	TUI_i
Being successfully interrupted: p_j starts speaking when p_i is speaking, p_i finishes his turn while p_j is still speaking.	BSI_i
Being unsuccessfully interrupted: p_j starts speaking when p_i is speaking, p_j finishes his turn while p_i is still speaking.	BUI_i
The total time that p_i speaks first after another speaker.	TSA_i
Ratio between the TSI_i and BSI_i.	$TSBI_i$
p_i floor grab: similar to TSI_i but if everybody in the group stops speaking.	SFG_i
Ratio between the total speaking length of p_i (no matter p_i speaks alone or at the same time with the others) to silence.	$FTSS_i$
Ratio between TSI_i and the total number of turns p_i has.	$RSIT_i$
Ratio between TUI_i and the total number of turns p_i has.	$RUIT_i$
Prosodic based NFs	**Abbreviation**
[Total, minimum, maximum, median, mean, standard deviation] speaking energy of p_i when there is no overlapping speech segment.	$TENoOver_i$ $MinENoOver_i$ $MaxENoOver_i$ $MedENoOver_i$ $MeanENoOver_i$ $StdENoOver_i$
[Total, minimum, maximum, median, mean, standard deviation] speaking energy of p_i.	TE_i $MinE_i$ $MaxE_i$ $MedE_i$ $MeanE_i$ $StdE_i$
[Minimum, maximum, median, mean, standard deviation] pitch for p_i when there is no overlapping speech segment.	$MinPNoOver_i$ $MaxPNoOver_i$ $MedPNoOver_i$ $MeanPNoOver_i$ $StdPNoOver_i$
[Minimum, maximum, median, mean, standard deviation] pitch for p_i.	$MinP_i$ $MaxP_i$ $MedP_i$ $MeanP_i$ $StdP_i$

B. Visual NFs

The visual NFs include *i)* visual focus of attention (VFOA) based features, *ii)* head activity based features (Head-Act) and *iii)* body activity based features (Body-Act). The descriptions of these features are given in Table VI with the corresponding abbreviations.

1) Visual Focus of Attention Based NFs: To obtain VFOA vector, the method proposed in [12], [17] was utilized. This method first finds facial landmarks per participant using the Constrained Local Model (CLM) [69] which converts the facial landmarks in 2D to 3D to detect the head pose representation (pan, tilt and roll). Then, the head pose representation is used to find VFOA. In this study, VFOA was defined as left, right, front and no-one such that if the participant is looking at the participant on his/her left, right, front or not looking at any other participant but somewhere else, respectively. The VFOA of a participant was modeled and estimated using SVM where the radial basis kernel function (RBF) with varying kernel parameters was used. In this study, on average 359.4 frames (labeled as left, right, front and no-one) per participant were used for SVM training. Lastly, the obtained VFOA was smoothed (the span used for the moving average was taken as 5) for de-noising [12], [17].

This VFOA detection method [12], [17] was tested on a subset which includes randomly chosen 23000 labeled frames (on average 359 per participant) and the accuracies 0.86, 0.85, 0.70, 0.50 were obtained for right, left, front and no-one, respectively. Using VFOA vector per one participant, in total eight features were extracted (Table VI).

2) Head Activity Based NFs: The Head-Act detection was adapted from [2]. Unlike [2], to detect and track the faces, we used the Viola-Jones [70] face detection algorithm which is based on Haar-like features and AdaBoost. A trained face detector which detects the faces with a rectangle bounding box was used. Using 25600 randomly selected frames (400 frames for each participant) 90% accuracy was obtained. After detecting the faces, the optical flow vectors (Lucas-Kanade algorithm [71]) of two consecutive frames within the face area were found. The optical flow vectors were used to define the average head motion per participant in 2D (resulted in 2 real-valued vectors). These vectors were binarized using a threshold which is the sum of the mean and the standard deviation of head motion per dimensions to distinguish the significant and not-significant head activities. After thresholding, two binary vectors (one for significant and zero for insignificant Head-Act such as small movements, noise, etc.) were combined with an OR operation to obtain a final binary Head-Act vector.

3) Body Activity Based NFs: The Body-Act detection was performed as given in [2], [24]. According to that algorithm [2], [24], first, image differencing was applied to detect the foreground pixels i.e. pixels belong to the participant. All foreground pixels except the head area were considered as a part of the body. Once, the difference image between two consecutive frames was obtained using a threshold (taken as 30), the moving and not-moving pixels were differentiated. So, if a pixels value was greater than the threshold used, that pixel was labeled as a moving pixel, otherwise it was labeled as a not-moving pixel. After finding the moving pixels, the total number of moving pixels in each frame was normalized by the size of the frame which resulted in a real-valued vector. Later, this vector was binarized using another threshold (taken as 5%) to differentiate the significant and insignificant body activities.

Using the obtained real-valued head (in total two) and body activity (in total one) vectors and the binary head and body activity vectors (one vector for each), for each participant, in total five features from head activity and four features from body activity were extracted (Table VI).

C. Audio-Visual NFs

Given that this section introduces audio-visual NFs, one can be interested in the synchronization of audio and video, al-

TABLE VI: Visual NFs for participant p_i.

VFOA based NFs	**Abbreviation**
The total time that p_i is being watched by the other participants.	TW_i
The total time that p_i is mutually looking at any other participants (mutual engagement (ME)).	TME_i
The total time that p_i is being watched by any other participants while there is no ME.	$TWerNoME_i$
The total time that p_i looks at other participants.	TL_i
The total time that p_i initiates the MEs with any other participants.	$TInitME_i$
For p_i the total time intercurrent between the initiation of ME with any other participants.	$TIntCME_i$
The total time that p_i is looking at any other participants while there is no ME.	$TLNoME_i$
Ratio between the TW_i and TL_i.	$RTWTL_i$
Head/body activity based NFs	**Abbreviation**
The total time that the head/body of p_i is moving.	THL_i, TBL_i
The total number of head/body activity turns for p_i where each turn represents a continuous head/body activity.	THT_i, TBT_i
Average head/body activity turn duration for p_i.	AHT_i, ABT_i
Standard deviation ($std.$) of head activity in x, y dimensions and the $std.$ of body activity for p_i.	$stdHx_i$, $stdHy_i$, $stdB_i$

TABLE VII: Audio-visual NFs for participant p_i.

VFOA w/speaking activity NFs	**Abbreviation**
The total time that p_i is looking at other participants while speaking.	$TLwS_i$
The total time that p_i is looking at other participants while not speaking.	$TLwNS_i$
The total time that p_i is being watched by the other (at least one) participants while speaking.	$TWwS_i$
The total time that p_i is being watched by 2 persons while speaking.	$T2WwS_i$
The total time that p_i is being watched by 3 persons (in our case by everybody) while speaking.	$T3WwS_i$
Ratio between $TLwS_i$ and looking at others while someone else is speaking.	$RTLwSTLwSS_i$
Ratio between $TLwS_i$ and $TLwNS_i$.	$RTLwSTLsNS_i$
The total time ME happens (with anyone) while p_i is speaking.	$TMEwS_i$
VA w/speaking activity NFs	**Abbreviation**
The total time that p_i has a visual activity while speaking.	$TVAwS_i$
The total number of visual activity turns (short turns are included) of p_i while speaking.	$THATwS_i$
The total number of visual activity turns without short movements (each turn lasts at least two seconds) when p_i is speaking.	$THTwS_i$
Average visual activity turn duration (short turns are also included) for p_i while speaking.	$AVAATwS_i$
The total visual activity interruptions while p_i is speaking: p_i has a visual activity when p_j is active and p_i become inactive before p_i but all these happens while p_i is speaking.	$TVAIwS_i$
The total visual activity floor grabs for p_i while speaking.	$TVAFGwS_i$

though all feature extraction steps applied after meetings were synchronized. All synchronization processes were performed manually by localizing the synchronization point that carried out by switching the room's lights on and off for the video and meanwhile emitting a chirp for the audio. There are some frame drops in audio and/or video but in the worst case the time difference between video and the corresponding audio is 0.78 seconds (approximately 16 frames). Hence, for the video or audio data whichever has a missing frame(s), during features extraction step, a spline interpolation was applied to recover the features corresponding to the missing frame(s).

The audio-visual features include *i)* VFOA with Speaking-Act based features (referred as VFOA-Spk-Act) and *ii)* VA with Speaking-Act based features (referred as VA-Spk-Act).

1) NFs Based on VFOA with Speaking-Act: After synchronization of audio and video, more specifically, after the Speaking-Act vectors (see Section IV-A1) and the VFOA vectors (see Section IV-B1) per participant were synchronized, the NFs (in total eight) defined in Table VII were extracted.

2) NFs Based on VA with Speaking-Act: The VA of a participant was found by applying an OR operation to binary Head-Act vector (Section IV-B2) and binary Body-Act vector (Section IV-B3). Therefore, in case there was a significant head and/or body activity the corresponding VA instance was set to one and in case there was no significant head and body activity, the corresponding VA instance was set to zero. The NFs extracted using VA with Speaking-Act are (in total six) given in Table VII with their abbreviations.

V. Prediction of Leadership Style (LS)

In this section, the methodology of Multiple Kernel Learning (MKL) is introduced briefly and Localized Multiple Kernel Learning (LMKL) [15], [16] i.e. the method we adapted for LS prediction is described with the reasonings why we are utilizing it. Finally, the other kernel based methods which were used for comparison are also summarized.

A. Multiple Kernel Learning

Multiple Kernel Learning (MKL) methods use a set of kernels in linear or non-linear way by automatically finding an optimal kernel combination from a large set of kernels. The feature combination and training are simultaneously performed. One advantage of MKL is that by using different kernels different feature subsets coming from multiple different sources, modalities and potentially having different notions of similarity can work together [15], [16]. Whereas the traditional way i.e. feature concatenation might cause overfitting especially on small training sets and also the specific statistical property of each feature vector can be ignored [72].

There have been extensive work on MKL in the literature. The simplest way to combine different kernels is using an un-weighted sum or product of kernels which give equal importance to all of them, although learning a weighted sum (or product) using a training set is a better strategy. Moreover, kernel combinations can be linear or nonlinear while nonlinear is less restrictive as compared to linear combination and can result in better classification results. For a comprehensive survey on different MKL methods and their comparisons, interested readers can refer to [16].

In this study, we utilized Localized Multiple Kernel Learning (LMKL) [15], [16]. LMKL uses nonlinear kernel weights combination such that different kernel weights are assigned to different regions of the feature space. This method consists of

two components: *i)* gating model which selects the optimum kernel function locally and *ii)* kernel based classifier. The optimization of these two components are not convex and are performed jointly with a two-step procedure such that: *i)* a gating model parameters are fixed to perform an optimization on kernel matrix and *ii)* the gating parameters are updated using the gradient found from the current optimization.

This method was preferred mainly because it allows to use the same type of kernel (e.g. linear, polynomial, Gaussian kernels) for different subset of data using a nonlinear gating model which eliminates the necessity of a grid search to find the best kernel function for each data subsets and due to its better prediction performance compared to many other MKL methods as shown in many papers such as [12], [15], [16].

LMKL can be combined with any kernel based learning algorithm. In this work, for fair comparisons it was combined with SVM which is the most popular learning method as shown in Table I. Therefore, the time complexity of LMKL is based on SVM and therefore it is quadratic programming [16]. It is required to solve a canonical SVM problem with the combined kernel obtained with the existing gating model and to calculate the gradients. As mentioned in [15], the gradient calculation step has ignorable time complexity compared to the SVM solver. As gating model, sigmoid and softmax functions were used with linear kernels varying from two to seven. As kernel parameter C (which is a trade-off parameter between model simplicity and classification error) was taken as 2^{i}, $i = -1, 1, 3...31$.

B. Competitive Learning Methods

The performance of LMKL was compared with single kernel SVM and another popular MKL method called Generalized Multiple Kernel Learning (GMKL) [73], [74]. SVM was applied using each feature group individually and when different feature groups were concatenated. For SVM, RBF and linear kernels were used. For GMKL, as the kernel based learning method SVM was used. Sum and product of RBF kernel subject to l_1 and l_2 regularizations were tested while the number of kernels was taken the same with the number of features (as applied in [12]). As kernel parameters C was taken as 2^{i}, $i = -1, 1, 3...31$ and RBF γ was used as 2^{j}, $j = -11, -9, -7...11$ for SVM and GMKL. GMKL was applied when different feature groups were concatenated.

VI. Experimental Work and Results

All the methods were applied for *a)* binary classification of autocratic versus democratic LS and *b)* multi-class classification of autocratic, democratic LS and not-a-leader using one-versus-one binary classifications (as suggested in [75] compared to applying one-versus-all). As mentioned before two different datasets: meeting segments and holistic meetings were used.

For training and testing, we applied leave-one-meeting-out and leave-one-meeting-segment-out cross validation approaches when meeting segments were used. We present the leave-one-meeting-out results only since the similar conclusions were acquired when leave-one-meeting-segment-out was applied although the performance of methods were better. For the experiments performed using the holistic meetings, leave-one-out cross-validation approach was used.

The performances of each method are expressed in terms of the detection rates (Eq. 1) and the geometric mean of detection rates (Eq. 2) as suggested to use by many studies such as [5] when there is a class imbalance like the data used have (in total 36 democratic, 13 autocratic and 147 not-a-leader exist for the meetings segments and 5 democratic, 7 autocratic and 36 not-a-leader exist for the holistic meetings when the videos having audio problems are discarded).

$$DetectionRate_c = \frac{\#CorrectlyPredictedSamples_c}{\#TotalSamples_c} \quad (1)$$

$$GeoMean = \Big(\prod_{1}^{N} DetectionRate_c\Big)^{1/N} \quad (2)$$

where c refers to class which can be autocratic, democratic and the not-a-leader in our case. N is the total number of classes which is equal to two for binary classification and three for multi-class classification.

The results given in Tables VIII, IX, X, and XI correspond to the best scores of each algorithm when audio NFs, visual NFs, audio-visual NFs and fusion of audio-video based NFs, respectively were used. With these analysis the contributions of different modalities and different feature groups were examined deeply. The shown results correspond to the average performances over cross validation folds and are the results with the highest GeoMean. In these tables, the overall best results for each metric are emphasized in bold-face while democratic LS detection rates are shown as DEM, autocratic LS detection rates are shown as AUT and not-a-leader detection rates are shown as NL.

For the evaluation of NFs including audio (i.e. Sections VI-A, VI-C and VI-D), 196 samples (4×49; each meeting segment has 4 participants and there are 49 meeting segments without any audio or annotation problems) from the dataset composed of meeting segments and 48 samples (4×12; each meeting has 4 participants and there are 12 meetings without any audio problems) from the dataset composed of holistic meetings were utilized. For the other experiments (Section VI-B), in total 264 samples (4×66; each meeting segment has 4 participants and there are 66 meeting segments in total) from the meeting segments and 64 samples (4×16; each meeting has 4 participants and there are 16 meetings in total) from the whole meetings were used for the evaluation of NFs.

A. Prediction of LS using Audio NFs

In Table VIII, the best results of each method for multi-class classification and binary classification using two datasets with audio NFs were given. SVM was applied using each feature group individually (shown as Speaking-Act-SVM, and Prosodic-SVM) and when they were concatenated (shown as All-SVM). GMKL and LMKL were applied using all features together (shown as All-GMKL and All-LMKL, respectively).

TABLE VIII: The best result of each method using audio NFs applied to meeting segments and meetings as a whole for multi-class and binary classifications. The best results are emphasized in bold-face. DEM is for democratic LS, AUT is for autocratic LS and NL is for not-a-leader.

Meeting Segments \Meetings as a Whole				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**
Speaking-Act-SVM	**0.73**\0.67	0.59\0.80	**0.70**\0.43	**0.94**\0.89
Prosodic-SVM	0.62\0.53	0.48\0.40	0.54\0.43	0.93\0.87
All-SVM	0.59\0.54	0.48\0.60	0.47\0.29	0.91\0.89
All-GMKL	0.61\0.75	0.53\0.60	0.47**0.72**	0.92\0.98
All-LMKL	0.69**0.86**	**0.62****1.00**	0.58\0.65	0.91**0.99**
Binary	**GeoMean**	**DEM**	**AUT**	
Speaking-Act-SVM	**0.83**\0.76	**0.89**\0.80	0.77\0.72	
Prosodic-SVM	0.69\0.66	**0.89**\0.60	0.54\0.72	
All-SVM	0.80\0.68	0.84\0.80	0.77\0.58	
All-GMKL	0.70\0.66	0.62\0.43	**0.78****1.00**	
All-LMKL	0.74**0.83**	0.70**0.86**	**0.78**\0.80	

TABLE IX: The best result of each method using visual NFs applied to meeting segments and meetings as a whole for multi-class and binary classifications. The best results are emphasized in bold-face. DEM is for democratic LS, AUT is for autocratic LS and NL is for not-a-leader.

Meeting Segments \Meetings as a Whole				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**
VFOA-SVM	0.44**0.79**	0.39**0.63**	0.27**0.88**	0.82\0.90
Head-Act-SVM	0.40\0.40	0.56\0.38	0.37\0.25	0.32\0.67
Body-Act-SVM	0.41\0.55	0.32**0.63**	0.27\0.38	0.82\0.69
All-SVM	0.58\0.43	0.54\0.38	0.43\0.25	0.85\0.84
All-GMKL	0.54\0.68	0.47\0.50	0.37\0.63	**0.92****0.98**
All-LMKL	**0.69**\0.76	**0.60****0.63**	**0.74**\0.75	0.75\0.94
Binary	**GeoMean**	**DEM**	**AUT**	
VFOA-SVM	0.65**0.87**	0.54**0.75**	0.79**1.00**	
Head-Act-SVM	0.70\0.69	0.71**0.75**	0.69\0.63	
Body-Act-SVM	0.70\0.75	0.71**0.75**	0.69\0.75	
All-SVM	0.64\0.50	**0.77**\0.50	0.53\0.50	
All-GMKL	0.58\0.81	0.43**0.75**	0.77\0.88	
All-LMKL	**0.72**\0.81	0.64**0.75**	**0.81**\0.88	

As seen in Table VIII, although LMKL performed the best (GeoMean) using the dataset composed of meetings as a whole, when the meeting segments were utilized, SVM performed much better than LMKL, especially with Speaking-Act features. Overall, GMKL performed worse than LMKL except autocratic LS detection of some cases. The effectiveness of Speaking-Act features with SVM was also validated in Section VI-F such that the best audio based NFs were detected as belong to that sub-group.

B. Prediction of LS using Visual NFs

The best results of each method for multi-class classification and binary classification using two datasets with visual NFs are given in Table IX. SVM was applied using each feature group individually (shown as VFOA-SVM, Head-Act-SVM, and Body-Act-SVM) and when they were concatenated (shown as All-SVM). GMKL and LMKL were applied using all features together (shown as All-GMKL and All-LMKL, respectively).

The results in Table IX shows that LMKL performed much better than SVM and GMKL when visual NFs extracted from meeting segments were used. For meeting segments, the LMKL's detection rates for autocratic LS were better than its detection rates for democratic LS. GMKL performed significantly better (using GeoMean, p-value < 0.01) than All-SVM for holistic meetings but performed worse than All-SVM for meeting segments. Besides, GMKL preformed worse than LMKL except binary classification of holistic data. When holistic meetings were used, VFOA based NFs performed much better than head/body-Act features and their fusion. This is perhaps because more training data was used to model the VFOA per person which resulted in better VFOA estimations as compared to modeling VFOA per person for each meeting segments independently. Furthermore, VFOA requires supervision while extraction of other features are unsupervised. As shown in Section VI-F, the majority of the visual NFs selected as the best features by LMKL were also VFOA-based and using only these selected features resulted in improved performance.

TABLE X: The best result of each method using audio-visual NFs applied to meeting segments and meetings as a whole for multi-class and binary classifications. The best results are emphasized in bold-face. DEM is for democratic LS, AUT is for autocratic LS and NL is for not-a-leader.

Meeting Segments \Meetings as a Whole				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**
VFOA-Spk-Act-SVM	0.55\0.60	0.56\0.40	0.31\0.58	0.95\0.92
VA-Spk-Act-SVM	0.49\0.48	0.42\0.40	0.31\0.29	0.92\0.98
All-SVM	0.54\0.55	0.45**0.60**	0.39\0.29	0.91\0.95
All-GMKL	0.45\0.61	**0.62**\0.40	0.16\0.58	0.94**0.99**
All-LMKL	**0.69****0.72**	**0.62****0.60**	**0.54****0.72**	**0.98**\0.87
Binary	**GeoMean**	**DEM**	**AUT**	
VFOA-Spk-Act-SVM	0.72\0.76	**0.95**\0.80	0.54\0.72	
VA-Spk-Act-SVM	0.59\0.59	0.73\0.40	0.47\0.86	
All-SVM	0.55\0.66	0.78\0.60	0.39\0.72	
All-GMKL	0.39**0.85**	0.24**0.72**	0.62**1.00**	
All-LMKL	**0.79**\0.68	0.85\0.58	**0.73**\0.80	

C. Prediction of LS using Audio-Visual NFs

For the evaluation of audio-visual NFs, SVM was applied using each feature group individually (shown as VFOA-Spk-Act-SVM and VA-Spk-Act-SVM) and when they were concatenated (shown as All-SVM). GMKL and LMKL were applied using all features together (shown as All-GMKL and All-LMKL, respectively).

As seen in Table X, LMKL performed the best using audio-visual NFs except binary classification of holistic data which GMKL performed significantly better (p-value < 0.01) than all others, for GeoMean and particularly for the detection of autocratic leaders.

D. Prediction of LS using Audio and Video Based Features Together

The audio NFs (Section IV-A) and visual NFs (Section IV-B) were used together as well. In supplementary material, visualizations of the meeting segments and holistic meetings in terms of these features after applying t-SNE [76] to represent them in 2-dimensional feature space are given.

TABLE XI: The best result of each method using audio and visual (Vis-Aud) NFs together applied to meeting segments and meetings as a whole for multi-class and binary classifications. The best results are emphasized in bold-face. DEM is for democratic LS, AUT is for autocratic LS and NL is for not-a-leader.

Meeting Segments \Meetings as a Whole				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**
Vis-Aud-All-SVM	0.61\0.54	0.50\0.60	0.47\0.29	**0.95**\0.92
Vis-Aud-All-GMKL	0.52\0.78	0.56\0.60	0.27**0.79**	**0.95****0.99**
Vis-Aud-All-LMKL	**0.69****0.83**	**0.59****0.80**	**0.58****0.79**	**0.95**\0.92
Binary	**GeoMean**	**DEM**	**AUT**	
Vis-Aud-All-SVM	0.66\0.68	0.64\0.80	0.47\0.58	
Vis-Aud-All-GMKL	0.63\0.76	0.47\0.58	0.84**1.00**	
Vis-Aud-All-LMKL	**0.84****0.83**	**0.77****0.86**	**0.92**\0.80	

TABLE XII: The best result of SVM and LMKL using correlated NFs. The best results are emphasized in bold-face. DEM is for democratic LS, AUT is for autocratic LS and NL is for not-a-leader.

Meetings as a Whole				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**
Corr-Speaking-Act-SVM	**0.60**	**0.40**	**0.58**	0.92
Corr-Prosodic-SVM	0.46	**0.40**	0.29	0.84
Corr-All-SVM	0.53	**0.40**	0.43	0.87
Corr-All-GMKL	0.52	**0.40**	0.36	**0.95**
Corr-All-LMKL	0.58	**0.40**	**0.58**	0.85
Binary	**GeoMean**	**DEM**	**AUT**	
Corr-Speaking-Act-SVM	0.83	0.80	**0.86**	
Corr-Prosodic-SVM	0.41	0.40	0.43	
Corr-All-SVM	0.66	0.60	0.72	
Corr-All-GMKL	0.68	0.58	0.80	
Corr-All-LMKL	**0.83**	**0.86**	0.80	

Using these NFs together, the best results for multi-class classification and binary classification for two datasets are given in Table XI which are shown as All-SVM, All-GMKL and All-LMKL for prediction using SVM, GMKL and LMKL, respectively.

As seen in Table XI, LMKL performed much better than SVM and GMKL when audio-based and video-based NFs used together. The only exception was the autocratic leader detection when binary classification was applied to holistic meetings where GMKL performed significantly better (p-value < 0.01) than all others. The successful performances of LMKL were perhaps thanks to the ability of defining the optimum kernel weights for different feature subsets that are coming from different sources and having different notions of similarity.

Results in Table XI are also comparable with the results shown in Table VIII (audio NFs) and Table X (audio-visual NFs) given that they utilized exactly the same training and testing sets per cross validation folds. Such a comparison showed that LMKL's results with fusion of audio and visual NFs were *i)* the same or much better than LMKL's audio NFs results except multi-class classification of holistic meetings, *ii)* much better than LMKL's audio-visual NFs results except multi-class classification of meeting segments such that they both performed the same. In summary, for the majority of the experiments, using fusion of audio and visual NFs improved the overall prediction performance (GeoMean) and also the democratic, autocratic detection rates individually.

Overall, LMKL outperformed all other learning methods. However, it is worth to mention that for some cases (i.e. democratic detection vs. autocratic detection, meeting segments vs. holistic meetings and binary classifications vs. multi-class classification, while different NFs were used), GMKL were competitive with LMKL. The most frequent case that GMKL performed better (or as good as) than LMKL was "the detection of autocratic leaders for binary classification of holistic meetings". However, even for this combination, its GeoMean i.e. the overall detection performance of GMKL was never better than LMKL (except Table X; binary classification of whole meetings) which means that although GMKL was good at detecting autocratic leaders, it was not good enough to detect the democratic leaders. This is perhaps because GMKL, which is a generalization method (i.e., unlike LMKL, GMKL defines the kernels without considering the local properties/similarities of the data), tended to be biased towards the classification of majority class (that is autocratic leaders for holistic meetings). Whereas, for LMKL such a bias was less obvious, therefore this resulted in better democratic detection rate and GeoMean for LMKL.

E. Correlation Analysis

The correlation analysis was performed using the data belonging to all participants and the data belonging to only leaders. Since, SYMLOG results correspond to holistic meetings only, the evaluation was performed using holistic data. The correlation analysis was performed between results derived from SYMLOG friendliness and each NF using Pearson's Correlation Coefficient [77] such that the correlation coefficient of two random variables (in our case, SYMLOG-friendliness scores of all participants and leaders only versus the corresponding NF, one per each time) is a measure of their linear dependence. The correlation coefficients found with the corresponding significance and the definition of low-high-medium correlations are given in supplementary material.

The results showed that 18.4% of the NFs were not correlated, while 50% were low correlated but without significance, and 15.44% were significantly medium or high correlated. Additionally, 14.71% of the NFs were medium/high correlated without significance which means that the amount of data used was not enough to make a conclusion for those features' correlations. The NFs found as significantly high/medium correlated were: TSI, SFG, RSIT, TE, MedE, MeanE, StdE, TENoOver, MedENoOver, MeanENoOver, StdENoOver, MinP, and MinPNoOver such that all of them are based on audio.

One can assume that correlated features can perform better prediction compared to using all features. To investigate this, SVM, GMKL and LMKL were applied using correlated features. In Table XII the corresponding best results (for multi-class classification and binary classification) using the whole meetings are given. During this analysis, the features which presented a correlation (low, medium, high) but without a significance were not included.

When the results using correlated NFs (Table XII) and all NFs (Table VIII) were compared, it was seen that there was a decrease in prediction performance particularly for multi-class classification. For binary classification, better performance was obtained for GMKL and for Speaking-Act-SVM when they were applied to correlated features while the performance of LMKL stayed the same and for all other cases the performances dropped. Here, it is worth mentioning that a correlation can only indicate the presence or absence of a relationship but not the nature of the relationship. Hence, correlation is not a causation and there is always the possibility that another variable(s) can influence the results [78]. Hence, it is possible that correlated features could perform worse classification performance as opposed to using all features (as also shown in [17]).

Motivating from these conclusions, we propose using a simple heuristic (Section VI-F) to identify the features which can perform better prediction in contrast to using all features. This heuristic is based on LMKL which finds the optimum kernel weights that reflect the relative contribution of each feature to classification task [15], [16].

F. Prediction of LS using Best Features Only

Although, in general, LMKL performed well for prediction of LS, in fact, the results obtained using both meeting segments and whole meetings showed that there were some redundant and irrelevant features which misguided the classification. On the other hand, the correlation analysis results showed that only Speaking-Act and prosodic features were correlated (even though correlation is not a classification implication). One can find all these results foreseeable given that we proposed these features motivated from other studies which were related to LS prediction.

In this section, we present a simple heuristic to select the best features which potentially can perform better prediction results compared to using features all together. Regarding this, given that important features have higher combination weights, the average absolute kernel weights per NFs obtained from LMKL were used. By using meeting segments and whole meetings, we extracted the kernel weights from training of LMKL which corresponds to the analysis shown in Section VI. Once the kernel weights were obtained they were scaled to zero-one interval and a threshold which is sum of mean and standard deviation of normalized kernel weights were used to identify the important (best) features. In detail, features having values greater than or equal to the found threshold were selected as best features.

SVM was applied to the selected best features and the results were compared with baseline results i.e. when SVM applied to all features. The corresponding results are given in Table XIII for audio NFs only, visual NFs only, audio-visual NFs and fusion of audio and visual NFs, respectively.

Using selected audio NFs, the overall prediction results (GeoMean), democratic and autocratic detection results were all improved compared to using all audio NFs, as can be seen in Table XIII. In total 18 out of 37 audio NFs were selected while the majority of them were based on Speaking-Act. These NFs were: TMSL, TSSL, FMSSL, TST, TSI, TUI, TSA, TSBI, FTSS, RSIT, RUIT, MeanE, StdE, MeanENoOver, StdENoOver, MinP, MedP and MedPNoOver.

TABLE XIII: The comparison of SVM results using all and selected NFs per feature groups with the total number of features used ($\#Fea$). DEM is for democratic LS, AUT is for autocratic LS and NL is for not-a-leader.

Audio NFs:	**Meeting Segments \Meetings as a Whole**				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**	**#Fea**
All-SVM	0.59**0.54**	0.48**0.60**	0.47**0.29**	**0.91**\0.89	37
Selected-SVM	**0.66****0.54**	**0.59****0.60**	**0.54****0.29**	**0.91****0.92**	18
Binary	**GeoMean**	**DEM**	**AUT**		**#Fea**
All-SVM	0.80\0.68	0.84\0.80	**0.77**\0.58		37
Selected-SVM	**0.82****0.93**	**0.87****1.00**	**0.77****0.86**		18

Video NFs:	**Meeting Segments \Meetings as a Whole**				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**	**#Fea**
All-SVM	**0.58**\0.43	0.54\0.38	**0.43**\0.25	**0.85**\0.84	17
Selected-SVM	0.52**0.66**	**0.64****0.50**	**0.43****0.63**	0.50**0.92**	6
Binary	**GeoMean**	**DEM**	**AUT**		**#Fea**
All-SVM	0.64\0.50	0.77\0.50	**0.53**\0.50		17
Selected-SVM	**0.66****0.63**	**0.81****0.63**	**0.53****0.63**		6

Audio-Visual NFs:	**Meeting Segments \Meetings as a Whole**				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**	**#Fea**
All-SVM	**0.54****0.55**	0.45**0.60**	**0.39**\0.29	0.91**0.95**	14
Selected-SVM	0.50\0.53	**0.56**\0.40	0.24**0.43**	**0.95**\0.87	7
Binary	**GeoMean**	**DEM**	**AUT**		**#Fea**
All-SVM	0.55\0.66	0.78\0.60	**0.39**\0.72		14
Selected-SVM	**0.60****0.83**	**0.92****0.80**	**0.39****0.86**		7

Audio and Video NFs:	**Meeting Segments \Meetings as a Whole**				
Multi-Class	**GeoMean**	**DEM**	**AUT**	**NL**	**#Fea**
All-SVM	0.61\0.54	**0.50****0.60**	0.47\0.29	**0.95**\0.92	54
Selected-SVM	**0.66****0.63**	**0.50****0.60**	**0.62****0.43**	0.91**0.98**	24
Binary	**GeoMean**	**DEM**	**AUT**		**#Fea**
All-SVM	0.66\0.68	0.64**0.80**	0.47\0.58		54
Selected-SVM	**0.73****0.83**	**0.87****0.80**	**0.62****0.86**		24

As seen in Table XIII, using selected video NFs, for all cases and for all metrics, better results were obtained except multi-class classification of meeting segments. For multi-class classification of meeting segments, although the democratic detection rate improved significantly and autocratic detection rate stayed the same, since the not-a-leader detection rate was significantly worse, the overall performance (GeoMean) was also worse. In total 6 out of 17 visual NFs were selected while majority of them were based on VFOA and no feature was selected from body activity. These NFs were: TW, TME, TWerNoME, TInitME, TLNoME, and stdHy.

By using selected audio-visual NFs better or the same performances for all metrics were obtained for all binary classifications (Table XIII). However, for multi-class classifications using all audio-visual NFs performed better than selected audio-visual NFs while either democratic or autocratic detection rate was better (but not both at the same time). In total 7 out of 14 visual NFs were selected while 6 out of 7 features were from VFOA with Speaking-Act. These NFs were: TLwS, TLsNS, TWwS, T3WwS, RTLwSTLsNS, TMEwS, and TVAwS.

Using the 18 selected audio NFs and the six selected visual NFs together, the overall prediction results, democratic and autocratic detection results were all improved as compared to using all 54 features (see Table XIII for quantitative results).

To sum up, only in three out of 32 cases, the classification

performance (GeoMean) of selected features were worse than the classification performance of all features. One, four, and five times out of 32, the GeoMean, the democratic and the autocratic detection rate were found the same. But considering the numbers of features selected which were much less than the total numbers of features, these results still present an improvement.

VII. Conclusions and Future Work

In this study, a comprehensive survey about leadership style identification and related research performed by SSP and social psychology literatures were presented. Motivating from the effectiveness of nonverbal features (NFs) presented in reviewed papers, a framework for the prediction of leadership styles of an emergent leader was proposed. This approach was applied to an in-the-wild dataset having two different annotations i.e. for meeting segments and whole meetings. For the first time, in this study, localized multiple kernel learning method (LMKL) and various NFs extracted from different modalities were utilized for the detection of leadership styles. By using LMKL (which generally showed better prediction results as compared to other kernel based methods) and a simple heuristic proposed, the best performing features were selected. Using selected features only resulted in improved (or similar) prediction rates for any type of NFs, using any type of data and any classification approach (multi-class or binary) which is a significant outcome given that much less features were used. When fusion of audio and video-based features were used, the effectiveness of LMKL was more explicit given that it performed the best overall. These results confirmed the ability of LMKL to exploit different modalities and various types of features having different notion of similarity which lead to a more effective classification performance. We also analyzed the correlation between NFs extracted and the results obtained from social psychology questionnaire. High correlations were obtained for audio-based NFs. In conclusion, it was detected that some features were better for discriminating the autocratic and democratic leadership styles. For instance, *i)* speaking activity as compared to prosodic, *ii)* VFOA and VFOA with speaking activity as opposed to the other visual and audio-visual NFs, respectively, and *iii)* the fusion of audio and visual NFs as compared to audio only and audio-visual NFs performed better.

A limitation of this work could be seen as the size of the data used, although the proposed method tested on two different types of data and to the best of our knowledge, there is no other in-the-wild dataset for leadership style prediction. Still, as future work, further investigation by adding additional meetings will be carried out. Moreover, the proposed framework will be tested on different settings such as *i)* for more crowded scenarios by applying some transfer learning and domain adaptation methods which could utilize the learnt model and best performing NFs of this study and *ii)* for meetings having acquainted participants (note that in this study meetings composed of unacquainted participants were used to avoid possible confounds based on familiarity). Another future work is to predict the leadership style by modeling the interactions between group members as sequences of events using audio, video and audio-visual NFs, while the co-occurrences of NFs can also be investigated. Furthermore, ensemble of LMKL, motivated by the effective performance of LMKL shown in this work (such that it performed the best in general which was enough to show the goodness of the NFs used without the necessity to resort to more elaborated algorithms), will be tested after increasing the size of the dataset.

Acknowledgments

The authors are grateful to all participants who joined the group tasks to constitute the dataset and all annotators who worked on annotation of this dataset.

References

[1] S. Feese, A. Muaremi, B. Arnrich, G. Trster, B. Meyer, and K. Jonas, "Discriminating individually considerate and authoritarian leaders by speech activity cues." IEEE SocialCom/PASSAT, 2011, pp. 1460–1465.
[2] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.
[3] D. Sanchez-Cortes, "Computational methods for audio-visual analysis of emergent leadership," *PhD Thesis, EPFL, Lausanne*, 2013.
[4] T. Nanjundeswaraswamy and D. Swamy, "Leadership styles," *Advances in Management*, vol. 7, no. 2, pp. 57–62, 2014.
[5] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653–1672, 2015.
[6] S. Feese, B. Arnrich, G. Trster, B. Meyer, and K. Jonas, "Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion." IEEE SocialCom/PASSAT, 2012, pp. 520–525.
[7] A. Pierro, L. Mannetti, E. De-Grada, and A. W. Kruglanski, "Autocracy bias in informal groups under need for closure," *Personality and Social Psychology Bulletin*, vol. 29, no. 3, pp. 405–417, 2003.
[8] B. Bass, *The Bass Handbook of Leadership: Theory, Research and Managerial Applications*. Free Press, 2008.
[9] D. Foster, "A method of comparing follower satisfaction with the authoritarian, democratic, and laissez-faire styles of leadership," *Communication Teacher*, vol. 16, no. 2, pp. 4–6, 2002.
[10] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal Communication in Human Interaction*. Boston: 8th Edition, Wadsworth, Cengage Learning, 2013.
[11] O. Aran and D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in small group conversations," in *ICPR*, 2010, pp. 3687–3690.
[12] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Identification of emergent leaders in a meeting scenario using multiple kernel learning." ACM ICMI-ASSP4MI, 2016, pp. 3–10.
[13] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations from nonverbal activity cues," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 3, pp. 501–513, 2009.
[14] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1018–1031, 2014.
[15] M. Gonen and E. Alpaydin, "Localized multiple kernel learning," in *ICML*, 2008, pp. 352–359.
[16] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.
[17] C. Beyan, N. Carissimi, F. Capozzi, S. Vascon, M. Bustreo, A. Pierro, C. Becchio, and V. Murino, "Detecting emergent leader in a meeting environment using nonverbal visual features only." ACM ICMI, 2016, pp. 317–324.
[18] B. Bass and E. Riggio, *Transformational Leadership*. Psychology Press, 2006.
[19] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "Identifying emergent leadership in small groups using nonverbal communicative cues." ACM ICMI-MLMI, 2010, pp. 8–10.

[20] S. Feese, B. Arnrich, G. Troster, B. Meyer, and K. Jonas, "Detecting posture mirroring in social interactions with wearable sensors." Int. Symp. on Wearable Computers, 2011, pp. 119–120.

[21] D. B. Jayagopi and D. Gatica-Perez, "Mining group nonverbal conversational patterns using probabilistic topic models," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 790–802, 2010.

[22] O. Aran and D. Gatica-Perez, *Gevers T, Salah AA (eds), Computer analysis of human behavior, Analysis of group conversations: modeling social verticality*, London, Ed. Springer, 2011.

[23] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "Detecting emergent leaders in small groups using nonverbal behavior," *Technical report, Idiap Research Institute*, 2011.

[24] O. Aran and D. Gatica-Perez, "One of a kind: inferring personality impressions in meetings," in *ACM ICMI*, 2013, pp. 9–13.

[25] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *ACM ICMI*, 2015, pp. 15–22.

[26] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez, "Linking speaking and looking behavior patterns with group composition, perception, and performance." ACM ICMI, 2012, pp. 433–440.

[27] D. B. Jayagopi and D. Gatica-Perez, "Discovering group nonverbal conversational patterns with topics," in *ACM ICMI-MLMI*, 2009, pp. 3–6.

[28] O. Aran and D. Gatica-Perez, "Estimating dominance in small group meetings with audio-visual fusion of nonverbal cues," *Technical report, Idiap Research Institute*, 2010.

[29] D. Sanchez-Cortes, D. Jayagopi, and D. Gatica-Perez, "Predicting remote versus collocated group interactions using nonverbal cues." Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing, in ACM ICMI-MLMI, 2009.

[30] C. Jie and P. Peng, "Recognize the most dominant person in multi-party meetings using nontraditional features," in *IEEE ICIS*, 2010, pp. 312–316.

[31] H. Hung and D. G. Perez, "Identifying dominant people in meetings from audio-visual sensors," in *IEEE FG*, 2008, pp. 1–6.

[32] D. Sanchez-Cortes, O. Aran, D. B. Jayagopi, M. S. Mast, and D. Gatica-Perez., "Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition," *Journal on Multimodal User Interfaces*, vol. 7, no. 1–2, pp. 39–53, 2012.

[33] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Predicting the dominant clique in group conversations with nonverbal cues," in *ACM-MM*, 2008, pp. 15–22.

[34] D. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *ACM ICMI*, 2008, pp. 45–52.

[35] U. Avci and O. Aran, "Predicting the performance in decision-making tasks: From individual cues to group interaction," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 643–658, 2016.

[36] ——, "Effect of nonverbal behavioral patterns on the performance of small groups." Workshop on Understanding Modeling Multiparty Multimodal Interactions, 2014, pp. 9–14.

[37] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, 2005.

[38] D. B. Jayagopi, B. Raducanu, and D. Gatica-Perez, "Characterizing conversational group dynamics using nonverbal behaviour." IEEE ICME, 2009, pp. 370–373.

[39] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe, "On the relationship between head pose social attention and personality prediction for unstructured and dynamic group interactions." ACM ICMI, 2013, pp. 3–10.

[40] S. Escalera, O. Pujol, P. Radeva, J. Vitria, and M. Anguera, "Automatic detection of dominance and expected interest," in *EURASIP Journal on Advances in Signal Processing*, 2010, pp. 1–12.

[41] S. Escalera, R. Martinez, J. Vitria, P. Radeva, and M. Anguera, "Dominance detection in faceto-face conversations," in *IEEE CVPR Workshops*, 2009, pp. 856–861.

[42] R. Rienks and D. Heylen, "Dominance detection in meetings using easily obtainable features," *Lecture Notes in Computer Science, Springer, Berlin*, vol. 3869, pp. 76–86, 2006.

[43] G. Chittaranjan, O. Aran, and D. Gatica-Perez, "Exploiting observers' judgements for nonverbal group interaction analysis," in *IEEE FG*, 2001, pp. 734–739.

[44] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *ACM ICMI*, 2008, pp. 53–60.

[45] D. Jayagopi, T. Kim, A. Pentland, and D. Gatica-Perez, "Privacy-sensitive recognition of group conversational context with sociometers," *Springer Multimedia Systems*, vol. 18, pp. 3–14, 2012.

[46] F. Valente, S. Kim, and P. Motlicek, "Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus," in *INTERSPEECH*, 2012, pp. 1183–1186.

[47] K. Kalimeri, B. Lepri, T. Kim, F. Pianesi, and A. Pentland, "Automatic modeling of dominance effects using granger causality," *Human Behavior Understanding, Lecture Notes in Computer Science, A. Salah, Lepri, B., Eds., Springer, Berlin/Heidelberg*, vol. 7065, pp. 124–133, 2011.

[48] F. Kalimeri, B. Lepri, O. Aran, D. Jayagopi, D. Gatica-Perez, and F. Pianesi, "Modeling dominance effects on nonverbal behaviors using granger causality," in *ACM ICMI*, 2012, pp. 23–26.

[49] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro, "Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection," in *Workshop Tagging Mining and Retrieval of Human Related Activity Information*, 2007, pp. 9–14.

[50] H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez, "Associating audio-visual activity cues in a dominance estimation framework." IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2008, pp. 2160–7508.

[51] D. B. Jayagopi, "Computational modeling of face-to-face social interaction using nonverbal behavioral cues," *Master Thesis, EPFL, Lausanne*, 2011.

[52] H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Investigating automatic dominance estimation in groups from visual attention and speaking activity," in *ACM ICMI*, 2008, pp. 233–236.

[53] J. Bullee, "Detection of leadership in informal (small) groups based on cctv information," *Master Thesis, The Univeristy of Twente, The Netherlands*, 2013.

[54] J. A. Hall, L. S. LeBeau, and E. J. Coats, "Nonverbal behavior and the vertical dimension of social relations: A meta-analysis," *Psychological Bulletin*, vol. 131, no. 6, pp. 898–924, 2005.

[55] R. T. Stein, "Identifying emergent leaders from verbal and nonverbal communications," *Personal. Social Psychol.*, vol. 32, no. 1, pp. 125–135, 1975.

[56] M. S. Mast, "Dominance as expressed and inferred through speaking time: A meta-analysis." *Human Communication Research*, vol. 28, no. 3, pp. 420–450, 2002.

[57] J. F. Dovidio and S. L. Ellyson, "Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening." *Social Psychology Quartely*, vol. 45, no. 2, pp. 106–113, 1982.

[58] C. Anderson and G. J. Kilduff, "Why do dominant personalities attain influence in face-to-face groups? the competence-signaling effects of trait dominance," *Journal of Personality and Social Psychology*, vol. 96, no. 2, pp. 491–503, 2009.

[59] J. E. Baird, "Some non-verbal elements of leadership emergence," *Southern Speech Communication Journal*, vol. 42, no. 4, pp. 352–361, 1977.

[60] A. K. Kalma, L. Visser, and A. Peeters., "Sociable and aggressive dominance: Personality differences in leadership style?" *Leadership Quarterly*, vol. 4, no. 1, pp. 45–64, 1993.

[61] D. Johnson and F. Johnson, *Joining together: Group theory and group skills*. Prentice-Hall, Inc., 1991.

[62] R. Bales, *SYMLOG: case study kit with instructions for a group self study*. New York: The Free Press, 1980.

[63] R. Koenigs, *SYMLOG reliability and validity*, San Diego: SYMLOG Consulting Group, 1999.

[64] A. Hare, R. Polley, and P. Stone, *The Symlog Practitioner: Applications of Small Group Research*. New York: Praeger Press, 1998.

[65] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 847–860, 2011.

[66] S. Gonzalez and M. Brookes, "PEFAC- a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[67] J. Markel, "The sift algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, pp. 367–377, 1972.

[68] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered HMMs," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 509–520, 2006.

[69] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *BMVC*, 2006, pp. 929–938.
[70] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE CVPR*, 2001, pp. 511–518.
[71] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt, "Performance of optical flow techniques," in *IEEE CVPR*, 1992, pp. 236–242.
[72] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *CoRR*, vol. abs/1304.5634, 2013.
[73] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *ICCV*, 2007.
[74] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *ICML*, 2009.
[75] C. Hsu and C. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, pp. 415–425, 2002.
[76] V. der Maaten Laurens and G. Hinton, "Visualizing data using t-sne." *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[77] K. Pearson, "Notes on regression and inheritance in the case of two parents." Proceedings of the Royal Society of London, 1895, pp. 240–242.
[78] J. Aldrich, "Correlations genuine and spurious in pearson and yule," *Statistical Science*, vol. 10, no. 4, pp. 364–376, 1995.

Cigdem Beyan received the BEng. degree in Computer Engineering from Baskent University, Ankara, Turkey in 2008, MSc. degree in Informatics (Information Systems) from Middle East Technical University, Ankara, Turkey in 2010 and Ph.D. degree in Informatics (Computer Vision) from Institute of Perception, Action and Behaviour in University of Edinburgh, Edinburgh, UK in 2015. She is currently a Postdoctoral Researcher at the Istituto Italiano di Tecnologia (IIT), Genoa, Italy in the department of Pattern Analysis and Computer Vision (PAVIS). Among her main research interest there are social signal processing, multimodal data analysis, behaviour understanding, trajectory analysis, anomaly detection, classification for imbalanced data and active learning. She is a member of IEEE and also an Associate Fellow of the Higher Education Academy in recognition of attainment against the UK Professional Standards Framework for teaching and learning support in higher education from 2014.

Francesca Capozzi is a postdoctoral fellow in the Department of Psychology at the McGill University (Montral, Canada). Her research interests include visual social attention and psychology of human social groups. More specifically, her work examines how natural social interactions modulate attention and gaze behavior in groups and multi-agent contexts.

Cristina Becchio is Senior Researcher at the Italian Institute of Technology (C'MON Unit), Genova, and Professor at the Psychology Department, University of Turin. After studying philosophy at the University of Turin, in 2001 she joined a PhD Program in Cognitive Science. In 2004 she moved on to a postdoctoral work on cognitive neuroscience and has since then been working on motor cognition. Her research focuses on the control and perception of complex body motion during human social interaction, specifically upon the ability to decode others' mental states from movement observation. In 2013 she started as Principal Investigator a five year project on "Intention-from-movement understanding" (I Move U) funded by the ERC, ERC-St-2012.

Vittorio Murino (SM'02) received the Laurea degree in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Genova, Genoa, Italy, in 1989 and 1993, respectively. He is a Full Professor with the University of Verona, Verona, Italy, and the Director of Pattern Analysis and Computer Vision (PAVIS) Department, Istituto Italiano di Tecnologia, Genoa. From 1995 to 1998, he was an Assistant Professor with the Department of Mathematics and Computer Science, University of Udine, Udine, Italy. Since 1998, he has been with the University of Verona. He was the Chairman of the Department of Computer Science from 2001 to 2007, where he was the Coordinator of the Ph.D. program in Computer Science from 1999 to 2003. He is scientific responsible of several national and European projects, and an Evaluator of EU project proposals related to several frameworks and programs. He is currently with the Istituto Italiano di Tecnologia, leading PAVIS Department involved in computer vision, machine learning, and image analysis activities. He has coauthored over 400 papers published in refereed journals and international conferences. His current research interests include computer vision, pattern recognition, and machine learning, more specifically, statistical and probabilistic techniques for image and video processing, with applications on (human) behavior analysis and related applications such as video surveillance, biomedical imaging, and bioinformatics. Prof. Murino is a member of the technical committees of most significant computer vision and pattern recognition conferences and a Guest Co-Editor of special issues in relevant scientific journals. He is also an Editorial Board Member of Computer Vision and Image Understanding, Machine Vision and Applications, and Pattern Analysis and Applications journals. He has been an IAPR Fellow since 2006.

"This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes more information regarding the dataset utilized, the nonverbal features (NFs) used and reports the results obtained from the correlation analysis between the extracted NFs and the results of SYMLOG questionnaire (friendliness sub-scale). Contact cigdem.beyan@iit.it for further questions about this work."