A Sequential Data Analysis Approach to Detect Emergent Leaders in Small Groups

Cigdem Beyan, *Member, IEEE,* Vasiliki-Maria Katsageorgiou, and Vittorio Murino, *Senior Member, IEEE*

***Abstract*—This paper addresses the problem of predicting emergent leaders (ELs) in small groups i.e. meetings. This is a long-lasting research problem for social and organizational psychology and a relevant problem that recently gained a momentum in social computing. Towards this goal, we propose a novel method, which analyzes the temporal dependencies of the audio-visual data by applying unsupervised deep learning generative models (feature learning). To the best of our knowledge, this is the first attempt that sequential data processing is performed for EL detection. Feature learning results in a single feature vector per a given time interval and all feature vectors representing a participant are aggregated using novel fusion techniques. Lastly, the emergent leader detection is performed using the state of the art single and multiple kernel learning algorithms. The proposed method shows (significantly) improved results as compared to the state of the art methods and it can be adapted to analyze various small group interactions given that it is a general approach.**

***Index Terms*—Emergent leader, social signal processing, non-verbal features, small group interactions, sequential data analysis, temporal data, Restricted Boltzmann Machines.**

I. Introduction

DETECTION of emergent leaders (ELs) using nonverbal signals (such as gaze, speaking activity, facial expressions, body activity, etc.) has been studied since mid-seventies by social psychologists [1]. However, automatic realization of EL detection recently became a popular research topic among social signal processing (SSP) community. In SSP, identification of ELs was particularly investigated for small-group interactions i.e. in meetings where participants were previously unacquainted and the leaders naturally appeared by showing some leadership characteristics such as authority, dominance, influence and control over other participants [2]–[6].

The conventional way of automatic identification of ELs in small group meetings typically contains *i)* extraction of nonverbal features (NFs), and *ii)* application of various computational methods for learning and prediction [2]–[6].

NFs are mainly extracted from two different modalities, i.e. *a)* audio (e.g. the total speaking length, the total speaking turn duration, and the total amount of successful interruptions [5], [6]), and *b)* video (e.g. the total time that head/body is moving, the total number of head/body activity turn, and the total time of being gazed [2], [3]). Additionally, the fusion of these two modalities was frequently applied e.g. in [4], [7].

In general, NFs can be extracted in two ways; ignoring the sequential aspect of data while features extracted are representing the whole meeting (or a meeting segment) such as extracting the total amount of existence of an attribute. Alternatively, one can consider the time dimension of data, i.e. performing sequential data analysis such that the features extracted are in terms of a certain time window (e.g. per each video/audio frame) and are also dependent to previous and/or next observations. To the best of our knowledge, all EL detection studies in SSP utilized non-sequential NFs (which are referred as static NFs in the rest of this paper) although as also mentioned in [8], examining social interactions with sequential data analysis can be more promising.

In this paper, we propose processing the audio-visual data by capturing the sequential aspects of it to detect the ELs in small group meetings. We use NFs extracted for each video frame, which is synchronized with the audio. Then, various unsupervised generative methods, which are able to represent the data by modeling the temporal relationship of it, are applied (namely, feature learning using sequential data). The results of this feature learning step are fused using novel techniques resulting in a compact feature vector that represents the nonverbal behavior of a given participant in a meeting segment. Once the compact feature representations are obtained for each participant, they are given as input to classifiers; Support Vector Machines (SVM) and Localized Multiple Kernel Learning (LMKL) [9], [10] (since they are the state of the art methods for EL detection) for the prediction of the ELs.

The main contribution of this work is presenting a novel approach for the detection of ELs in small group meetings, which shows improved results as compared to state of the art methods. Additionally,

- this is the first time that detection of ELs is investigated using sequential data, which is realized using unsupervised feature learning methods that have never been utilized for this problem,
- novel techniques (based on bag of words, and covariance) to fuse the results of the unsupervised feature learning step, resulting in compact feature representations, and significantly improving the detection rates of ELs and not-ELs in small group meetings as compared to traditional fusion approaches are proposed,
- last but not the least, a comprehensive review regarding pairs of the NFs and the computational methods applied with their corresponding results are presented. These results can be

C. Beyan and V. Murino are with Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, 16152, Italy, V.-M. Katsageorgiou is with Advanced Robotics, Istituto Italiano di Tecnologia (IIT), Genoa, 16163, Italy and V. Murino is also with Department of Computer Science, University of Verona, Verona, Italy (e-mail: cigdem.beyan@iit.it; vasiliki.katsageorgiou@iit.it; vittorio.murino@iit.it).

Manuscript received xx xx, XX; revised xx xx, xx.

utilized as the baselines of the dataset used in this study, which can be useful/helpful for future studies using the same dataset.

The obtained results shows that the claims *i)* sequential data analysis for EL detection results in better performance than using static NFs, and *ii)* using bag of words and covariance-based fusion methods perform better than traditional fusion techniques are correct. These claims are made and proved first time in this study specifically for EL detection, but, it is important to highlight that the proposed pipeline can be easily adapted for the analysis of other small group social interactions which might result in improved performances for them too.

Additionally, the proposed method has some advantages as compared to the state of the art methods i.e. the studies using static NFs for EL detection. The state of the art works generally included extracting statistical properties such as the maximum, the minimum, the standard deviation of the total amount of existence of a static NF (as applied in [2], [3], [6]), and/or the skewness, the number of zero crossings, the number of mean crossings etc. of a NF (as applied in [4]). Therefore, NFs used in these studies were hand-crafted and statistical properties were usually applied with a trial and error procedure while justifying the reason of their effective performance is also harder. On the other hand, in this study, we are not investigating any specific statistical property of a NF, but instead the unsupervised feature learning methods applied realize this automatically, while the resulting NFs are data-driven [11], [12].

The results of unsupervised feature learning step (a single feature vector per frame, representing a participant) can be combined in different ways as well. For instance, feature vectors can be used with the same label of the corresponding participant for learning and classification or majority voting can be applied to determine the final predicted label of that participant (in this study, we referred them as late fusion, for the detailed definitions of late fusion methods, see Section VI). These approaches are simpler than the proposed fusion algorithms, but our proposals (Section IV-C) show significantly better results as compared to them.

The rest of this paper is organized as follows. The previous studies about EL detection and an extensive review on papers analyzing small group interactions with temporal data are given in Section II. The dataset utilized and its annotation procedure are described in Section III. In Section IV, the NFs with their extraction methods, and the components of the proposed method: unsupervised feature learning, fusion of feature learning results, learning/prediction are described. The details of experimental analysis, the brief description of the comparative methods and comparative NFs are given in V. Following that, the results of the proposed method, the comparative methods and the previously reported best methods are compared in Section VI. Finally, the paper is concluded with discussions and future work in Section VII.

II. Related Work

In this section, first, we briefly reviewed the SSP studies, which examined automatic EL detection in small group meetings. Particularly, the NFs, and the computational models they utilized were discussed. Then, we present a comprehensive survey of the SSP studies, which analyzed the data sequentially to investigate various small group interactions. For both parts, the main similarities/differences between previous works and our study are highlighted.

A. EL Detection in SSP

This review is limited to SSP studies. Interested readers can refer to [1], [13], which surveyed the emergence of leadership and nonverbal behaviors in social psychology literature.

The NFs used for EL detection in small group interactions can be grouped into three as: *i)* audio-based, *ii)* video-based and *iii)* audio-visual. Until now, as audio-based features; speaking activity (speech activity, speaking turn) based features [4]–[6] and prosodic features [6] were utilized. On the other hand, eye gaze, which was modeled in terms of visual focus of attention (VFOA) in [2], [3], head/body activity [3], [4], [6], and 2D-body pose [4] were used to extract video-based NFs. As audio-visual NFs, in [7], VFOA-based NFs, which were extracted depending on the speaking activity (such as looking while speaking, being looked while speaking, etc.) were used. Whereas, in [4], speaking activity-based NFs were fused with head/body activity and 2D-body pose-based NFs.

Overall, speaking activity and VFOA-based NFs performed better than any other NFs for automatic detection of ELs in small group meetings [2]–[4], [6]. 2D-body pose-based NFs showed better performance as compared to head/body activity-based NFs while their fusion with VFOA-based NFs performed better than any combination between speaking activity NFs, head/body activity NFs and 2D-body pose-based NFs. Motivating from findings of previous works, in this study, some of the VFOA-based NFs presented in [2], [3] and speaking activity i.e. speaking diarization results are utilized (see Section IV-A for more detail).

For automatic detection of ELs in meetings, different computational methods have been employed. As supervised learning method; Support Vector Machines (SVM) [2]–[4], [4], [6], [7] and as unsupervised learning method; rank-level fusion approach (RLFA) [5]–[7] were the most popular techniques. For the first time, Localized Multiple Kernel Learning (LMKL) [9], [10] was applied for this problem in [3], which showed state of the art results when various video-based NFs were used. The better prediction performance of LMKL was validated in [4] for automatic identification of ELs when multimodal NFs used and in [13] for the prediction of leadership style of ELs in small group meetings when audio-based, video-based and audio-visual NFs were utilized.

As mentioned in Section I, different from any EL detection studies in the SSP literature, we approach this problem by analyzing the data sequentially, which is realized using deep learning techniques: Conditional Restricted Boltzmann Machines (CRBM) [14] and Recurrent Neural Network Restricted Boltzmann Machines (RNNRBM) [15], motivating from their effective performances for many different applications as shown in [14], [16]. Once sequential data is processed and results of this step are combined by applying novel fusion

techniques, we apply SVM and LMKL (as they are the state of the art classifiers for this problem and using them also allows us to make fairer comparisons with previous studies) for EL prediction.

B. Investigation of Small Group Interactions using Temporal Data

The review presented in this section includes social interactions, mainly, during small-group meetings and is limited to the human-human interactions. In terms of applications, the SSP studies presenting temporal data analysis can be grouped as follows: *i)* meeting segmentation based on individual and group actions [17]–[22], *ii)* detection of dominance effects [23], [24], *ii)* social role recognition [25]–[31], *iii)* detection of group interest-level [32], *iv)* detection of influence rankings [33], [34], *v)* performance prediction of a group in decision making tasks [35], and *vi)* detection of essential social interaction predicates [36].

The majority of these studies utilized NFs: audio-based (speech activity, energy, pitch, speaking rate) [17]–[22], [24]–[35], [37], [38], video-based (head activity, hand activity, body movements, gaze, head orientation, head and body gestures, etc.) [17]–[19], [21]–[24], [28], [31], [32], [34]–[38] while few of them [18], [21], [25], [26] included verbal features (n-grams, Linguistic Inquiry and Word Count, and occurrence of positive and negative words) as well.

All these studies analyzed small group meetings composed of 4-participants (which is in line with the corpus we utilize in this study), except [36], which is based on dyadic communication during games (i.e. not during meetings like others).

The most frequently applied sequential data analysis methods were Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN), which were applied to different applications (see Table I). As seen in Table I, Granger Causality was mainly applied for detection of dominance effects, while Influence Model (IM) and Conditional Random Fields (CRF) were applied for role recognition. Hierarchical Probabilistic Model used in [37] showed effective results for discovering interactions such as question-and-answer and addressing behavior followed by back-channel responses.

Regarding the sequential model used, the most similar study to ours is [36]. In [36], detection of essential social interaction predicates (i.e. joint attention and entertainment) from audio-visual dyadic interactions was addressed using a "tower game" dataset. Discriminative Conditional Restricted Boltzmann Machine (DCRBM) was used, which includes a discriminative property in addition to the generative property of CRBM, unlike our study, which applies SVM and LMKL as discriminative methods. The same authors [36] applied the proposed DCRBM for classifying sequential multimodal raw sensor data in [39], which can be applied to detect various small group interactions during meetings, although not yet applied for this purpose.

As can be seen from the review given in this section, temporal data has never been used for EL detection in small groups. Herein, we claim that applying unsupervised feature learning (UFL) using sequential data and the proposed fusion techniques, which to the best of our knowledge never applied with UFL, performs better than the state of the art techniques.

TABLE I: Sequential data analysis methods applied by SSP studies.

Methods: **References**
Hidden Markov Models: [17], [19]–[22], [27], [29], [32], [35]
Dynamic Bayesian Networks: [18], [20], [22], [30], [33], [34], [38]
Granger Causality: [23], [24]
Conditional Random Fields: [25], [26]
Hierarchical Probabilistic Model: [37]
Influence Model: [28], [31]
Discriminative Conditional Restricted Boltzmann Machine: [36]

III. The Dataset

For the evaluation of the proposed method, the leadership dataset used in [2]–[4], [13] was utilized. This dataset contains 16 meetings, which is in total 393 minutes and each meeting lasts from 12 to 30 minutes. In each meeting, there are same gender, similar age, unacquainted four-participants. The participants performed a winter or desert "survival" tasks [40], which are the most common tasks in small group decision making particularly for dominance and leadership detection (see [2] for details).

The data acquisition set up included five cameras. Four of them were frontal (resolution of 1280x1024 pixels and frame rate of 20 fps) i.e. capturing each participant individually and the last one (a standard camera, resolution of 1440x1080 pixels and frame rate of 25 fps) captured the whole scene, which was used for data annotation [2]. Audio (sample rate=16 kHz) was recorded with four wireless lapel microphones, each one connected to participant's corresponding frontal camera.

The annotations are for the meeting segments (in total 75 meeting segments), which were obtained by dividing the 16 meetings into small chunks (see [2] for the reason of such a video segmentation) and demonstrates "the most EL", "the least EL" and "the rest". The statistics corresponding to data annotation, i.e. the agreement between annotators and the reliability scores can be found in [2].

Four out of 16 meetings have audio problems, thus they were discarded for the experiments performed in this study, given that we are using audio-based NFs as well. In total, 58 meeting segments, which constitute 232 samples (4x58, each segment has four participants) were used for the experimental analysis. It is important to highlight that the audio problem was also considered in [13], but in that study [13], less samples were used as there were additional annotation problems regarding leadership style prediction, whereas there is no annotation problem regarding EL detection, which results in more samples to be used in this study.

IV. Proposed Method

The proposed method has four components. For each participant, NFs are first extracted from video and audio, which are based on visual focus of attention (VFOA) and speaking activity (SpeakAct). This step results in a feature matrix of $M \times F$ where F is equal to the number of frames in a meeting segment and M is equal to the number of NFs (which is 6 in

our case, see Section IV-A for details). Then, an unsupervised feature learning method (UFL, either CRBM or RNNRBM), which is suitable for sequential data processing (Section IV-B) is applied. As a result of this step, a matrix of $N \times H$ where N is equal to F for RNNRBM and is equal to F $minus$ time-delay (see Section IV-B1 for the definition) for CRBM and H is equal to the number of hidden units of RBM (see Section IV-B for the definition) is obtained. The results of UFL step are fused using either Bag of Words-based or Covariance-based method (Section IV-C), which results in a feature vector of length K, where K is equal to $(H \times (H+1)/2)$ for Covariance-based fusion and is equal to the number of unique binary configurations of the hidden units for the entire dataset for the Bag of Words-based method. Finally, the obtained high-level features ($1 \times K$ for each participant) are used for learning and classification (SVM or LMKL is utilized as they are the state of the art classifiers for ELship) where the classification results can be the most EL, the least EL or the rest. This proposed methodology is illustrated in Figure 1.

The motivations behind the choice of features and models used are as follows. Previous works [2]–[4], [6], showed that SpeakAct-based NFs and VFOA-based NFs perform better than any other audio or video-based NFs for automatic detection of ELs in small group meeting. Therefore, we use VFOA-based NFs and SpeakAct. For EL detection in meetings, it was shown in [4] that RBM could improve the results. Herein, we claim that performing EL detection with sequential data analysis can result in better performance than using static NFs, and given that two of the closest methods to RBM for modeling sequential data are CRBM and RNNRBM, we utilize these two models. Using them is also advantageous as they do not require large amount of training data particularly as compared to many other deep models. Still, the UFL step can be altered to other sequential models. A discussion regarding that can be found in Section VII. Bag of words and covariance have been used as feature representations for different applications and showed state of the art results. We adapt these methods by claiming that they can perform better than most popular fusion techniques (see Table III for comparisons). SVM and LMKL are used to be in line with the state of the art works of EL detection, i.e. [2]–[4], [6] and to provide fair comparisons.

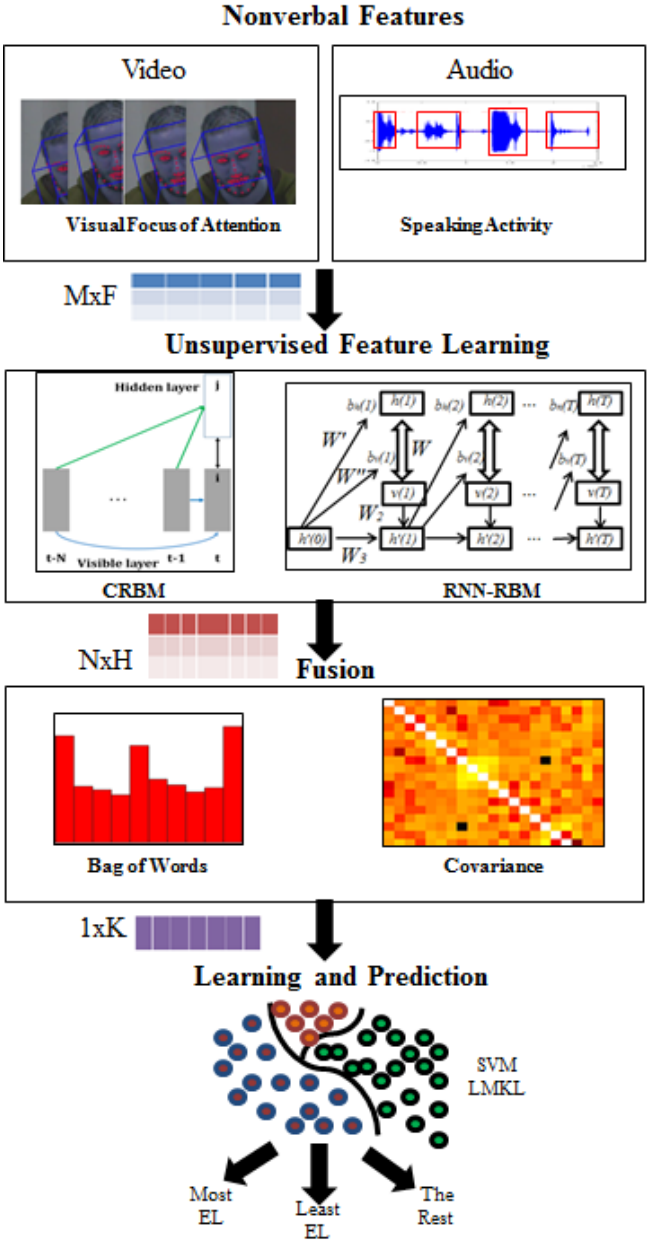

Fig. 1: The overview of the proposed approach.

A. Nonverbal Features and Their Extraction

Visual focus of attention (VFOA) is any location that a participant is looking at [41]. Assuming that there are four participants in a small group meeting, a participant's VFOA can be the person who is on the left, right, front side of him/her or elsewhere. VFOA can be inferred using the head pose (specifically pan and tilt). We adapt the VFOA detection algorithm presented in [2], which was applied for EL detection using the same dataset. Briefly, the facial landmarks using the Constrained Local Model (CLM) are found and facial landmarks in two-dimensions are converted to three-dimensions to detect the head pose [42]. A SVM model per participant are trained using few VFOA-labeled head poses and, then this model are used to find the VFOA for the rest of the video frames. Finally, the whole VFOA vector is de-noised using a moving-average filter with length of five.

Speaking activity (SpeakAct) is the result of speaker diarization process, which segments the input audio such that the audio is labeled in terms of "who speak when" [43]. In this work, we apply the same methodology applied in [4] and [13] for EL and ELship detection, respectively. This method is based on energy such that a speaking segment of a participant is detected if the energy difference between that participant's energy value and the mean value of the other participants is greater than a threshold. This results in a binary vector such that one represents "speaking" and zero represents "not-speaking". As the last step, this vector is de-noised by merging the segments of the participant within one second.

The binary (i.e. the existence of any feature is represented as one, while the non-existence is represented as zero) NFs extracted for a participant p_i at time t are as follows. In total six NFs are extracted while five of them are VFOA-based and the last one is the SpeakAct.

1) Looking at a participant when there is no mutual engagement (ME) with him/her.
2) Being looked by a participant while there is no ME with him/her.
3) Being looked by two participants while there is no ME with them.
4) Being looked by three participants while there is no ME with them.
5) ME with any participant.
6) Speaking.

B. Unsupervised Feature Learning using Sequential Data

Unsupervised feature learning (UFL) provides a joint representation of the input features by modeling the non-linear interactions between them. It is also task-independent as being unsupervised. Deep learning was widely explored for feature learning [11], [12]. For instance, [11], [12] demonstrated that data-driven learned features (i.e. deep learning-based features) performed better than the hand-crafted features. For EL detection in meetings, it was shown in [4] that UFL could improve the EL prediction results when it was applied as a postprocessing method while various types of NFs were the input. Different from [4] (the only study that applied UFL, particularly, Deep Boltzmann Machines (DBMs) i.e. the stacks of multiple RBMs, for EL detection), herein, we apply UFL methods suitable for processing sequential data.

As UFL methods *i)* Conditional Restricted Boltzmann Machines (CRBM) [14] and *ii)* Recurrent Neural Network Restricted Boltzmann Machines (RNNRBM) [15] are used.

1) Conditional Restricted Boltzmann Machines (CRBMs): CRBMs are non-linear generative models with real-valued Gaussian visible units (v), which can model also temporal dependences in the data [14]. This can be done by treating the visible variables in the previous time slice(s) as additional fixed inputs. To do so, two extra types of directed connections are added, which turn the RBM into a Conditional RBM: i) autoregressive connections from the past N configurations (time-steps) of the v to the current visible configuration and ii) connections from the past M configurations of the v to the current hidden configuration. The autoregressive weights can model linear, temporally local structure very well, leaving the hidden units (h) to model nonlinear, higher-level structure. N and M are tunable parameters and for simplicity, in this study, they are initialized with the same value, which is also called order of the model or time-delay.

As all RBM models, the CRBM defines a joint probability distribution over the visible units v_t and the hidden units h_t, with the difference that the hidden units are also conditioned on the past $v_{<t}$:

$$p(v_t, h_t|v_{<t}) = \frac{exp(-E(v_t, h_t|v_{<t}))}{Z(v_{<t})} \quad (1)$$

where, the partition function Z is constant with respect to v_t and h_t but depends on $v_{<t}$.

In this study, the parameters of CRBMs are set as follows: h is set equal to the number of v or 1.5, 2, 3, 3.5, 4, 5 times of that, while the batch size, the learning rate, and the time delay are set to 512, 0.001, and $\{5, 7\}$. CRBMs are trained for 10000-30000 epochs, till we obtain a reconstruction error of magnitude of 0.001. The best results were (given in Section VI) obtained when the number of h was set equal to the number of v and time-delay was taken as 5.

2) Recurrent Neural Network Restricted Boltzmann Machines (RNNRBM): The RNNRBM is an energy-based model for density estimation of temporal, high-dimensional sequences, like RBM. It is essentially a series of conditional RBM (one at each time step) whose parameters are determined by a deterministic RNN. In other words, the RNNRBM model extends the RNN by adding an RBM at each time step and uses the prediction capability of the two models to learn temporal dependencies of the data.

In more detail, during training, given the visible vector at time t, the RNN hidden state and the associated bias vectors at time t are deterministic and can be readily computed for each training sequence.

The implementation of the model in this work follows the original work i.e. [15]. In this study, the parameters of RNNRBMs are set as follows. The number of hidden layers of RNN and RBM are taken as equal to the number of the visible units or 2, 3 times of that, keeping the number of the RBM hidden units equal to the number of the recurrent hidden units. The batch size are set to 1024, the learning rate is set to 0.001 and the training is continued until convergence in the RBM's energy function is observed ($\sim$ 6000-7000 training epochs).

C. Fusion of Feature Learning Results

CRBM and RNNRBM result in a single feature representation for each frame. These representations are combined using the techniques based on *i)* Bag of Words (BWF) and *ii)* Covariance (CF) such that each participant in a meeting segment could be represented with a single, more compact and potentially more discriminative feature vector. To the best of our knowledge, this is the first time (not only for EL detection but in general) that the output of CRBM and RNNRBM are combined using these methodologies.

1) Bag of Words-based Fusion (BWF): After the training process of the CRBM and the RNNRBM, each input sample is represented by a binary configuration of the hidden units. In other words, real valued input data is mapped to binary representations. Using these binary configurations, the bag of words-based fusion approach is applied as follows. First, the unique binary configurations in the dataset are found. These unique configurations construct the bag of words. Then, the occurrence of each of these unique words/configurations is computed for each participant, hence a histogram over the observed unique configurations per participant is obtained. Each histogram is then normalized *a)* by the total number of frames that belong to the corresponding participant or *b)*

by taking its L2 norm, to obtain the final feature vector of each participant. These final feature vectors, whose length is equal to the number of the unique binary configurations, are used for the detection of ELs.

2) Covariance-based Fusion (CF): The covariance matrix can effectively fuse multiple features while the diagonal entries of it are composed of the variance of each feature and the non-diagonal entries are composed of the correlations. Covariance has been used as a generic feature representation for different applications such as object detection, recognition and tracking [44]–[46], pedestrian detection [46], face recognition [47], action recognition [48]–[50], etc. However, to the best of our knowledge, it has never been combined with any UFL before.

The covariance matrix of the participant p_i in a given meeting segment is:

$$C_{p_i} = \frac{1}{n-1}\sum_{k=1}^{n}(f_{p_i}(k) - \mu_{p_i})(f_{p_i}(k) - \mu_{p_i})^T \quad (2)$$

where μ_{p_i} is the mean of the feature vectors f obtained by applying UFL (CRBM or RNNRBM), and n is the total number of frames (feature vectors). f is the probabilities of hidden units (latent states, see Section IV-B for more information).

The covariance matrixes do not lie on Euclidean space. However, matrix logarithm operation can be used to map covariance matrices from the Riemannian manifold to the Euclidean space. To do that, singular value decomposition, which decomposes the covariance matrix is applied as follows.

$$C_{p_i} = U\Sigma U^T \quad (3)$$

where U $d \times d$ orthonormal matrix, and Σ is the square diagonal matrix with nonnegative real numbers, eigenvalues on the diagonal. Then, the new representation of the covariance matrix can be written as:

$$C_{p_i}^{(log)} = \log(C_{p_i}) = U\Sigma' U^T \quad (4)$$

where Σ' is the $d \times d$ square matrix with logarithmic values of the eigenvalues on the diagonal. The compact feature representation obtained as a result of this fusion step is the all entries on and above (below) the diagonal of the covariance matrix.

D. Learning and Prediction

Support Vector Machine (SVM, i.e. the most popular classifier for EL detection) and Localized Multiple Kernel Learning (LMKL, i.e. the state of the art classifier for EL detection as shown in [3], [4]) are utilized to detect the most ELs, the least ELs and the rest.

- *SVM*: The Radial Basis kernel Function (RBF) and linear kernel are used. As kernel parameters C is taken as 2^{i}, $i = -1, 1, 3...31$ and RBF γ is used as 2^{j}, $j = -11, -9, -7...11$ as applied in [4], [13] for EL detection using the same dataset.
- *LMKL* [9], [10]: It assigns different kernel weights to different regions of the feature space and it uses a nonlinear kernel weights combination. There are two components of it namely, the gating model, which selects the optimum kernel function locally and the kernel-based classifier (which is SVM in this paper as applied in [3], [4], [13] for EL detection). As gating model, sigmoid and softmax functions are 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) is taken as 2^{i}, $i = -1, 1, 3...31$.

V. Experimental Analysis

Following the works [2]–[4], all the methods were applied for multi-class classification of the most EL, the least EL and the rest classes using one-versus-one binary classifications. During these analysis leave-one-meeting-out cross validation was applied in order to be able to compare the results with the previous studies' results (e.g. [2], [4]).

As the evaluation metric, the class detection rate (Eq. 5) was used. The best set of parameters was determined by using the geometric mean (GeoMean) of detection rates of each class (Eq. 6), which corresponds to the highest GeoMean obtained for a specific classification method and features used. These metrics were used in [2]–[4] for EL detection performance evaluation as well.

$$DetectionRate_c = \frac{\#CorrectlyPredictedSamples_c}{\#TotalSamples_c} \quad (5)$$

$$GeoMean = \Big(\prod_{1}^{N} DetectionRate_c\Big)^{1/N} \quad (6)$$

where c refers to class, which can be the most EL, the least EL and the rest in our case. N is the total number of classes, which is equal to three.

For the experiments with unsupervised feature learning, the parameter set, in other words, the features learnt whichever performed the best for SVM, were used with LMKL.

A. Comparative UFL Methods

- *Gaussian RBMs (GRBMs)*: Gaussian RBMs are RBMs having real-valued Gaussian visible units (v) and stochastic binary (Bernoulli) hidden units (h). Any set of h makes a linear contribution to the mean of each v and the resulting conditional distribution over the v is a Gaussian distribution. We performed various experiments with different settings of training parameters of GRBMs. h was set equal to the number of v or 0.8, 1.25, 1.5, 10 times of that while the batch size, and the learning rate were set to 8, and 0.001, respectively. We observed that 1000 training epochs were enough to obtain a low reconstruction error (i.e. 0.001), while also the number of the observed latent states was not changing.
- *Mean-Covariance RBMs (MCRBMs)*: MCRBMs are a type of RBMs, which are also capable of feature learning from real-valued input data. Similar to all RBM models, they have a bipartite undirected graph structure. They can be seen as a combination of a GRBM and a covariance RBM (cRBM) [51], thus their hidden units are divided into two sets: *i)* mean units (h_m) that model the mean of the input

TABLE II: Comparative VFOA-based [2]–[4], [13] and SpeakAct-based [6], [13] NFs extracted for participant p_i.

VFOA-based NFs
1.1. The total time that p_i has a ME at any other participants.
1.2. The total time that p_i is being watched by any other participants while there is no ME.
1.3. The total time that p_i is looking at any other participants while there is no ME.
1.4. The total time that p_i is being looked by two participants simultaneously, while there is no ME.
1.5. The total time that p_i is being looked by three participants simultaneously, while there is no ME.
1.6. The ratio between the features 1.2 and 1.3.
SpeakAct-based NFs
2.1. The total speaking length of p_i when at least one other participant is also speaking.
2.2. The total speaking length of p_i when nobody is speaking.
2.3. The ratio between the features 2.1 and 2.2.
2.4. The total number of speaking turns of p_i without utterances (a turn lasts minimum 2 seconds).
2.5. The average speaking turn durations of p_i.
2.6. The total time that p_i speaks first after another participant speaks.
2.7. The 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.

elements and *ii)* covariance units (h_c) that represent the pairwise dependencies between the visible ones, modeling their covariance structure. Like in all RBM models, there are no connections between the variables within the layers, thus, the variables of a layer are independent of each other. In this study, MCRBMs were trained with different parameter settings. Specifically, h_c was set equal to the number of v or 0.8, 1.25, 1.5, 10 times of that. h_m was set equal to the number of v or 0.8, 1.25, 1.5, 2, 8 times of that while the batch size, and the learning rate were set to 8, and 0.01, respectively. All training parameters were randomly initialized to small values as suggested in [51]. We observed that setting the number of h_m units bigger than the number of h_c units in general performed better than the opposite condition, while the best result (see Section VI) was obtained when the number of h_m units was set as 1.5 times of the number of v and the number of h_c units was set as 0.8 times of v. Also in this case, we observed that 1000 training epochs were enough for the model to converge, resulting in a stable reconstruction error and number of unique latent states.

B. Comparative Nonverbal Features

In Table II, the VFOA-based and SpeakAct-based static NFs are listed. These NFs are used to compare with the proposed method while the corresponding results are given in Table IV. The reason of selecting these NFs as comparative NFs are their better performances for EL detection as shown with the ablation studies performed in [2]–[4], [6].

VI. Results

SVM and LMKL were applied to static NFs given in Section V-B with and without applying MCRBM and GRBM. For the proposed method, to show the contribution of the proposed fusion step, this step was excluded such that the features obtained for each frame by applying CRBM and RNNRBM were used for learning and prediction directly (called late-fusion). In this context, it was assumed that each frame had the same label with the corresponding participant in the given meeting segment and the classification evaluation was performed *i)* per frames (shown as lateFusion) and also *ii)* per participant by applying majority voting to the predicted labels of frames (shown as lateFusion-maj). These results are given in Table III.

TABLE III: The best result of each method. Bold-face is used to highlight the best of all results for each evaluation metric.

Methods	**GeoMean**	**Most EL**	**Least EL**	**The Rest**
SVM	0.72	0.78	0.59	0.82
LMKL	0.80	0.80	0.64	**1.00**
MCRBM-SVM	0.73	0.73	0.66	0.80
MCRBM-LMKL	0.78	0.73	0.66	**1.00**
GRBM-SVM	0.72	0.80	0.59	0.78
GRBM-LMKL	0.82	0.80	0.69	**1.00**
CRBM-BWF-SVM (proposed)	0.72	0.76	0.73	0.69
CRBM-CF-SVM (proposed)	0.76	0.82	0.71	0.75
CRBM-BWF-LMKL (proposed)	0.80	0.73	0.71	**1.00**
CRBM-CF-LMKL (proposed)	**0.89**	**0.85**	**0.85**	0.98
RNNRBM-BWF-SVM (proposed)	0.72	0.71	0.73	0.71
RNNRBM-CF-SVM (proposed)	0.76	0.66	0.66	**1.00**
RNNRBM-BWL-LMKL (proposed)	0.85	0.78	0.80	0.98
RNNRBM-CF-LMKL (proposed)	0.83	0.75	0.75	**1.00**
CRBM-lateFusion-SVM	0.44	0.40	0.37	0.57
CRBM-lateFusion-LMKL	0.50	0.50	0.32	0.76
RNNRBM-lateFusion-SVM	0.25	0.49	0.07	0.44
RNNRBM-lateFusion-LMKL	0.50	0.38	0.44	0.74
CRBM-lateFusion-maj-SVM	0.52	0.35	0.68	0.60
CRBM-lateFusion-maj-LMKL	0.53	0.52	0.50	0.56
RNNRBM-lateFusion-maj-SVM	0.28	0.63	0.04	0.82
RNNRBM-lateFusion-maj-LMKL	0.52	0.46	0.38	0.81

Additionally, the results of the proposed method with various combinations (shown as CRBM-BWF-SVM, CRBM-CF-SVM, CRBM-BWF-LMKL, CRBM-CF-LMKL, RNNRBM-BWF-SVM, RNNRBM-CF-SVM, RNNRBM-BWL-LMKL, and RNNRBM-CF-LMKL) are also given in Table III to better understand the contribution of each component.

According to GeoMean results;

- the different combination of the proposed method performed (0.89, 0.85, 0.83) better than all other methods given in Table III,
- CRBM-CF-LMKL performed the best of all (0.89) and its performance was significantly better (paired t-test applied to GeoMean; p-value < 0.05) than SVM only, LMKL only, MCRBM with SVM and LMKL, and GRBM with SVM and LMKL,
- without any exception, all combinations of the proposed method performed significantly better (paired t-test applied to GeoMean; p-value < 0.01) than the experiments with lateFusion and lateFusion-maj,
- RNNRBM-BWL-LMKL performed significantly better (paired t-test applied to GeoMean; p-value < 0.05) than MCRBM-LMKL and any expriments with SVM given in Table III,
- any combination of the proposed method with LMKL performed better than any combination of it with SVM, which is in line with previous studies [2]–[4],
- the proposed method with Covariance-based Fusion (CF)

mostly (i.e. for CRBM-SVM, CRBM-LMKL, RNNRBM-SVM) performed better than the proposed method with Bag of Words-based Fusion (BWF), while for BWF, L2 normalization generally performed better than scale-normalization.

Concurrently, the best performing method to detect the most ELs and the least ELs was also CRBM-CF-LMKL; while its detection rate for "the rest" class was as good as LMKL, MCRBM-LMKL, GRBM-LMKL, CRBM-BWF-LMKL, RNNRBM-CF-SVM, RNNRBM-CF-LMKL, SpeakAct-Pose-LMKL (i.e. methods performing %100 detection rate).

A. Comparisons with the Previously Reported Best Methods

For the dataset [2] utilized in this study, previously reported best methods using different NFs are given in Table IV. All these NFs are static, as sequential data analysis have not been performed for the automatic detection of ELs in small group meetings before. It is important to highlight that, the VFOA-based and the SpeakAct-based NFs given in that table are not completely the same with the NFs given in Section V-B. This is because the NFs given in Section V-B are a subset that resulted in better prediction performances and, therefore, are chosen to compare with the proposed method. Additionally, all results given in Table IV were obtained using the data subset described in Section III, which is the same data subset used for the experiments given in Table III. Therefore, the results in Table IV can be slightly different from the previously reported results that can be found in [2]–[4].

In Table IV, RLFA stands for rank-level fusion approach (tests if any combination of NFs are performing better than a single NF) [6], HeadBodyAct refers to the head and body activity-based NFs (the total time that head/body is moving, the total time of head/body activity turns, the average of the head/body activity turn duration and the standard deviation of the head/body activity) [6], Pose represents the NFs based on some statistical measures of the angles from body joints that are obtained from 2-dimensional pose [4] and lastly, DBM refers to Deep Boltzmann Machines, which is composed of a MCRBM as the first layer and a Bernoulli-Bernoulli RBM as the second layer [4].

As seen in Table IV, the proposed method performed better (0.89, 0.85, 0.83, 0.80) than all other previously reported best performing methods in terms of GeoMean. Particularly, all combinations of the proposed method were significantly better (paired t-test applied to GeoMean; p-value < 0.05) than visual activity-based NFs: HeadBodyAct-SVM, HeadBodyAct-DBM-SVM, HeadBodyAct-DBM-LMKL, Pose-SVM, and Pose-DBM-SVM. Except CRBM-BWF-LMKL, RNNRBM-BWF-SVM, RNNRBM-CF-SVM and RNNRBM-CF-LMKL, the most EL detection rates of the proposed method were better than the most EL detection rate of the best of all method given in Table IV (i.e. VFOA-HeadBodyAct-LMKL), while all combinations of the proposed method performed better than VFOA-HeadBodyAct-LMKL for the least EL detection.

TABLE IV: Previously reported best performing methods applied to the same dataset [2] utilized in this study. The best results for each evaluation metric are emphasized in bold-face.

NFs-Classifier	GeoMean	Most EL	Least EL	The Rest
VFOA-RLFA [2], [3]	0.73	0.73	0.71	0.74
VFOA-SVM [2]–[4]	0.67	0.73	0.57	0.72
SpeakAct-SVM [4]	0.68	0.73	0.59	0.73
HeadBodyAct-SVM [4]	0.49	0.56	0.37	0.58
HeadBodyAct-LMKL [4]	0.65	0.61	0.47	0.97
HeadBodyAct-DBM-SVM [4]	0.45	0.54	0.57	0.30
HeadBodyAct-DBM-LMKL [4]	0.50	0.50	0.30	0.82
Pose-SVM [4]	0.50	0.45	0.49	0.57
Pose-LMKL [4]	0.64	0.61	0.45	0.97
Pose-DBM-SVM [4]	0.52	0.49	0.59	0.49
Pose-DBM-LMKL [4]	0.65	0.59	0.50	0.95
VFOA-HeadBodyAct-SVM [3], [4]	0.66	0.63	**0.75**	0.62
VFOA-HeadBodyAct-LMKL [3], [4]	**0.79**	**0.76**	0.64	**1.00**
SpeakAct-HeadBodyAct-SVM [4]	0.64	0.69	0.52	0.74
SpeakAct-HeadBodyAct-LMKL [4]	0.74	0.69	0.59	0.98
VFOA-Pose-SVM [4]	0.64	0.61	0.56	0.77
VFOA-Pose-LMKL [4]	0.78	0.75	0.64	**1.00**
SpeakAct-Pose-SVM [4]	0.65	0.71	0.52	0.76
SpeakAct-Pose-LMKL [4]	0.72	0.68	0.56	**1.00**

VII. Discussions and Future Work

In this study, for detection of ELs in small group meetings, we proposed a novel approach, which considers the temporal behavior of the audio-visual data and, thus, is unlike the traditional approach, which is utilizing non-sequential (static) NFs that are extracted from the whole meeting segments (or meetings) per participant and applying classifiers with these static NFs.

Deep learning-based generative feature learning methods, which were applied to low-level NFs extracted for each video frame (synchronized with audio frames), were able to represent the data by modeling the temporal relationship of it. The results of this stage were combined using proposed bag of words-based and covariance-based fusion methods, which resulted in a compact feature representation for a given participant in a meeting segment. This resulting feature vector was given as an input to the state of the art classifiers to detect the most ELs, the least ELs and the rest. The proposed method showed (significantly) improved results as compared to *i)* previously reported best methods applied to the same dataset, *ii)* the best performing static NFs with/without various feature learning methods, and *iii)* the traditional fusion techniques (late-fusion and late-fusion with majority voting).

In summary, this study showed an approach to analyze the sequential audio-visual data for EL detection in small groups. The proposed approach presented promising results such that it could be applied to different SSP problems especially to analyze small group multi-party environments, such as dominance detection, personal trait prediction, and rapport detection.

Although the results of the proposed method were already (significantly) better than the state of the art performances and, therefore, there was no necessity to resort to apply other algorithms, as future work, the following further investigations could be carried out. The unsupervised feature learning step could be exchanged with alternative methods such as discriminative CRBMs [36], [39], temporal RBMs [52], and recurrent temporal RBMs [53]. Realizing such an investigation could be more useful, once additional meetings are added to the dataset. Long Short-Term Memory Networks (LSTM) [54] could be

an alternative model to apply as well. However, it is highly possible that, an improved version of LSTM, which models long sequences might be needed to obtain better results than the results of the proposed method. Furthermore, other NFs such as visual activity-based NFs [4], [6] could be examined using the proposed pipeline.

REFERENCES

[1] D. Sanchez-Cortes, "Computational methods for audio-visual analysis of emergent leadership," *PhD Thesis, EPFL, Lausanne*, 2013.

[2] 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," in *Proc. of ACM ICMI*, 2016, pp. 317–324.

[3] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Identification of emergent leaders in a meeting scenario using multiple kernel learning," in *Proc. of ACM ICMI-ASSP4MI*, 2016, pp. 3–10.

[4] C. Beyan, V.-M. Katsageorgiou, and V. Murino, "Moving as a leader: Detecting emergent leadership in small groups using body pose," in *Proc. of ACM Multimedia*, 2017, pp. 1425–1433.

[5] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "Identifying emergent leadership in small groups using nonverbal communicative cues," in *Proc. of ACM ICMI-MLMI*, 2010, pp. 8–10.

[6] ——, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.

[7] 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.

[8] A. Vinciarelli and F. Valente, "Social signal processing: Understanding nonverbal communication in social interactions," in *Proc. of Measuring Behavior*, 2010, pp. 118–121.

[9] M. Gonen and E. Alpaydin, "Localized multiple kernel learning," in *Proc. of ICML*, 2008, pp. 352–359.

[10] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[11] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2011, pp. 215–223.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[13] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 441–456, 2018.

[14] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Two distributed-state models for generating high-dimensional time series," *Journal of Machine Learning Research*, vol. 12, pp. 1025–1068, 2011.

[15] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription," in *Proc. of ICML*, 2012, pp. 1159–1166.

[16] M. Langkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, pp. 11–24, 2014.

[17] 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.

[18] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic bayesian networks," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 25–36, 2007.

[19] 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.

[20] A. Dielmann and S. Renals, "Dynamic bayesian networks for meeting structuring," in *Proc. of IEEE ICASSP*, 2004, pp. 629–632.

[21] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interactions in meeting," in *Proc. of IEEE ICASSP*, 2003, pp. 748–751.

[22] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang, "Multimodal integration for meeting group action segmentation and recognition," *In: Renals S., Bengio S. (eds) Machine Learning for Multimodal Interaction. MLMI 2005. Lecture Notes in Computer Science*, vol. 3869, pp. 52–63, 2006.

[23] K. Kalimeri, B. Lepri, T. Kim, F. Pianesi, and A. S. 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.

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

[25] A. Sapru and H. Bourlard, "Automatic social role recognition in professional meetings using conditional random fields," in *Proc. of Interspeech*, 2013, pp. 1530–1534.

[26] ——, "Automatic recognition of emergent social roles in small group interactions," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 746–760, 2015.

[27] S. Favre, A. Dielmann, and A. Vinciarelli, "Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models," in *Proc. of ACM Multimedia*, 2009, pp. 585–588.

[28] W. Dong, B. Lepri, F. Pianesi, and A. Pentland, "Modeling functional roles dynamics in small group interactions," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 83–95, 2013.

[29] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli, "Role recognition in multiparty recordings using social affiliation networks and discrete distributions," in *Proc. of ACM ICMI*, 2008, pp. 29–36.

[30] A. Vinciarelli, F. Valente, S. H. Yella, and A. Sapru, "Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus," in *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, 2011, pp. 374–379.

[31] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *Proc. of ACM ICMI*, 2007, pp. 271–278.

[32] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. of IEEE ICASSP*, 2005, pp. 489–492.

[33] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and application of influence rankings in small group meetings," in *Proc. of ACM ICMI*, 2006, pp. 257–264.

[34] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase, "Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns," in *CHI Extended Abstracts on Human Factors in Computing Systems*, 2006, pp. 1175–1180.

[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] M. R. Amer, B. Siddiquie, A. Tamrakar, D. A. Salter, B. Lande, D. Mehri, and A. Divakaran, "Human social interaction modeling using temporal deep networks," *CoRR*, vol. abs/1505.02137, 2015. [Online]. Available: http://arxiv.org/abs/1505.02137

[37] K. Otsuka, H. Sawada, and J. Yamato, "Automatic Inference of Cross-Modal Nonverbal Interactions in Multiparty Conversations: who responds to whom, when, and how?" from gaze, head gestures, and utterances," in *Proc. of ACM ICMI*, 2007, pp. 255–262.

[38] K. Otsuka, Y. Takemae, and J. Yamato, "A probabilistic inference of multiparty-conversation structure based on markov switching models of gaze patterns, head directions, and utterances," in *Proc. of ACM ICMI*, 2005, pp. 191–198.

[39] M. R. Amer, T. Shields, B. Siddiquie, A. Tamrakar, A. Divakaran, and S. Chai, "Deep multimodal fusion: A hybrid approach," *International Journal of Computer Vision*, vol. 126, p. 440?456, 2017.

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

[41] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 928–938, 2002.

[42] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *IEEE ICCVW 300 Faces in-the-Wild Challenge*, 2013, pp. 354–361.

[43] 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.

[44] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. of ECCV*, 2006, pp. 589–600.

[45] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *Proc. of CVPR*, 2006, pp. 728–735.
[46] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on riemannian manifolds," in *Proc. of CVPR*, 2007, pp. 1–8.
[47] Y. Pang, Y. Yuan, and X. Li, "Gabor-based region covariance matrices for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 989–993, 2008.
[48] C. Yuan, W. Hu, X. Li, S. J. Maybank, and G. Luo, "Human action recognition under log-euclidean riemannian metric," in *Proc. of ACCV*, 2009, pp. 343–353.
[49] K. Guo, P. Ishwar, and J. Konrad, "Action recognition using sparse representation on covariance manifolds of optical flow," in *Proc. of AVSS*, 2010, pp. 188–195.
[50] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Proc. of IJCAI*, 2013, pp. 2466–2472.
[51] M. Ranzato and G. E. Hinton, "Modeling Pixel Means and Covariance Using Factorized Third-Order Boltzmann Machines," in *Proc. of CVPR*, 2010, pp. 2551–2558.
[52] I. Sutskever and G. Hinton, "Learning multilevel distributed representations for high-dimensional sequences," in *Proc. of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007, pp. 544–551.
[53] I. Sutskever, G. Hinton, and G. Taylor, "The recurrent temporal restricted boltzmann machine," in *Proc. of NIPS*, 2008, p. 1601?1608.
[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 9, pp. 1735–1780, Nov. 1997.

Cigdem Beyan received her MSc. degree in Informatics (Information Systems) from Middle East Technical University, Ankara, Turkey in 2010 and her 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, Genoa, Italy in the department of Pattern Analysis and Computer Vision. She has co-authored over 25 papers published in refereed journals and international conferences. Among her main research interest there are social signal processing, multimodal data analysis, human/animal behavior understanding, anomaly detection, deep learning and classification of imbalanced data. She is a reviewer of most significant multimedia, affective computing, computer vision and pattern recognition journals including multiple IEEE Transactions, and IEEE/ACM conferences. She is a Guest Co-Editor of a special issue in Frontiers in Robotics and AI and in the Editorial Board of ICES Journal of Marine Science covering area of applications of computer vision and machine learning in marine science. She is a member of IEEE and 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.

Vasiliki-Maria Katsageorgiou Vasiliki-Maria Katsageorgiou received her MSc. degree in Electrical and Computer Engineering from the Democritus University of Thrace, Xanthi, Greece in 2013 and her Ph.D. in Computer Vision and Machine Learning from the University of Genova, Genoa, Italy in 2017. She is currently a Postdoctoral Researcher at the department of Advanced Robotics (ADVR) of Istituto Italiano di Tecnologia (IIT), Genoa, Italy. Her research interests lie on the areas of Artificial Intelligence, Statistical Machine Learning and Computer Vision, with an emphasis on unsupervised data modeling, deep learning, probabilistic graphical models and time-series analysis.

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.