Detecting Emergent Leader in a Meeting Environment using Nonverbal Visual Features Only

Cigdem Beyan[1] Nicolò Carissimi[1] Francesca Capozzi[2] Sebastiano Vascon[3]
Matteo Bustreo[1] Antonio Pierro[4] Cristina Becchio[5,6] Vittorio Murino[1]
[1]Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy
[2]Department of Psychology, McGill University, Montreal, QC, Canada
[3]European Center for Living Technology, Ca' Foscari University of Venice, Italy
[4]Department of Social and Developmental Psychology, University of Rome La Sapienza, Italy
[5]Department of Psychology, University of Turin, Italy
[6]Robotics, Brain and Cognitive Sciences, Istituto Italiano di Tecnologia, Genova, Italy
(cigdem.beyan,nicolo.carissimi,cristina.becchio,matteo.bustreo,vittorio.murino)@iit.it
frcapozz@gmail.com,sebastiano.vascon@unive.it,antonio.pierro@uniroma1.it

ABSTRACT

In this paper, we propose an effective method for emergent leader detection in meeting environments which is based on nonverbal visual features. Identifying emergent leader is an important issue for organizations. It is also a well-investigated topic in social psychology while a relatively new problem in social signal processing (SSP). The effectiveness of nonverbal features have been shown by many previous SSP studies. In general, the nonverbal video-based features were not more effective compared to audio-based features although, their fusion generally improved the overall performance. However, in absence of audio sensors, the accurate detection of social interactions is still crucial. Motivating from that, we propose novel, automatically extracted, nonverbal features to identify the emergent leadership. The extracted nonverbal features were based on automatically estimated visual focus of attention which is based on head pose. The evaluation of the proposed method and the defined features were realized using a new dataset which is firstly introduced in this paper including its design, collection and annotation. The effectiveness of the features and the method were also compared with many state of the art features and methods.

CCS Concepts

•Information systems → Database management system engines;

Keywords

Emergent leadership; visual focus of attention; nonverbal features; social signal processing

1. INTRODUCTION

In the social context, a leader is a person who has authority and power over a group of people, who can exert his/her dominance, influence and control over them [33]. Besides, an emergent leader (EL) is a person who naturally shows these characteristics among a group [33]. Since leadership and coordination with colleagues are important for organizations, e.g. while hiring new managers, assessing someone's emergent leadership (ELship) skill is important. Therefore, automatic detection of leaders in ecological situations, e.g. in a meeting, when realized by social signal processing (SSP) techniques looking at nonverbal behavior only can be very helpful.

SSP is a relatively new research area which aims to understand and model the social interactions mainly considering the findings in social psychology [36]. This area covers many applications such as automatic inference of interactions [30], deception detection [28], detecting group interest level during meetings [14], modeling dominance in group conversations [2]. Furthermore, many SSP related corpus were issued such as [7] which covers head gestures, hand gestures, body movement and facial expressions that represent the nonverbal expressions.

While the verbal communication is a key in social interactions, it is known that a valuable amount of information is conveyed nonverbally [23]. Nonverbal communication includes cues like eye gaze, body gestures, facial expressions, etc. and has been studied by social psychology for a long time. The automatic extraction of nonverbal features from audio and/or video and determining the usefulness of each feature for a given SSP task have been investigated in many studies such as [2]. Specifically, gaze cues are very informative for social interactions and can be used to control the communication. It has been shown that when a speaker is addressing a person, the majority of the time he/she is gazing that person [22]. Given that eye gaze is the most reliable social attention cue [35], it should be effective in detecting ELs in meeting. This can be realized by the nonverbal features which are extracted from the visual focus of attention (VFOA) which is defined using gaze.

Automatically detecting ELs in a meeting environment is a new problem among other SSP topics which, to the best of our knowledge, was only investigated in [31] and in the

publications (such as [33, 32]) related to that work. For example, in [33], audio and video nonverbal features were used to detect the ELs. In that study, it has been shown that using nonverbal multi-modal features is more effective than using features extracted from single modality only, while visual nonverbal features generally were not performing as good as audio nonverbal features. Similarly, many works such as [20, 27] failed to show the additional value of video on the top of audio [2]. Many works (such as [18]) utilized only audio nonverbal features discarding the visual features completely. However, there are instances where audio is not available and the only way to analyze the social interaction is by using visual information which still should be performed precisely, as we show in the following.

Motivating from these arguments, in this study, we propose novel nonverbal visual features which are extracted from head pose of the people in a meeting environment to detect the EL. To this end, we present a new dataset which is publicly available[1]. The results show that, the new features are effective to detect the most and the least EL with 79% and 63% detection rate, respectively. The features' effectiveness were also validated with the social psychology questionnaires: *SYMLOG* [4, 24] and *GLIS* [25] (see Section 3.1) with a correlation analysis.

The contributions of this work can be listed as follows *i)* devising novel video based features using VFOA to detect the ELs in a meeting environment, whose correlation are also validated by social psychology questionnaires; *ii)* introducing a new dataset which can be used to detect EL during a meeting; and *iii)* presenting a comprehensive comparison among several VFOA methods involving imbalanced set classification methods which were never considered for the given SSP problem.

The rest of this paper is organized as follows. Section 2 discusses the previous studies related to ELship. The data collection is introduced in Section 3 including the annotation process and questionnaires that were applied. In Section 4 the proposed method which includes head pose estimation, VFOA detection and feature extraction are reported. We present the experimental results in Section 5. Finally, we conclude the paper with a discussion and future work in Section 6.

2. RELATED WORK

Here, we review ELship studies and the related concepts. For the interested readers, a detailed review on ELship in social psychology and social computing can be found in [31].

The SSP studies for detection of ELs in a meeting environment (including the dominance detection) can be divided into different categories in terms of: *i)* the nonverbal features that they utilized which were extracted from only audio (such as [12, 18]), from only video (the proposed method) and with the fusion of audio and video (such as [33, 2, 20]), *ii)* the type of visual data processing such as extracting head and body activity [33], or extracting VFOA from head pose [19, 31], *iii)* the evaluation of the method which can be detecting only EL (dominant) [31], or detecting the most and the least ELs (dominant) [2, 20].

In [33], ELship was investigated using nonverbal audio and video based features. The ELship were measured using the concepts dominance, influence, leadership and control. The main assumption of that study [33] was that a socially dominant person receives more frequent and longer lasting glances by the people, looks at others while speaking, uses more gestures, is more talkative and has longer turns. Using this assumption, nonverbal audio features such as: average speaking turn duration, total speaking turn, etc. and nonverbal visual features which were extracted from head and body activities were defined. Unlike our study, head pose was not used. In [12] the leadership styles: authoritarian and individually considerate were estimated using nonverbal features extracted from audio. The prediction of leadership style was performed using logistic regression using only the features obtained from leaders, meaning that not all participants which can be seen as a drawback. Another study which only uses audio to detect the dominant person in a meeting environment is [18]. In that study [18], speaker diarization was used and the results showed that dominance estimation is robust to increase in diarization noise. In [2], and [20], dominance in group conversation were investigated using short meeting segments (similar to ours). Different than our work, in [2], scenario meetings which contains assigned roles to each participant were also utilized. In both studies [2, 20], the most dominant and the least dominant people were identified independently (in contrast to these, we classified the most and the least EL with a common model) using different annotator agreements such that full agreement and majority agreement. In that studies [2, 20], body motions were detected to extract the nonverbal features from visual activity different than using head pose. As supervised learning method, a Gaussian Mixture Model (GMM) based ranking procedure using ranked Support Vector Machine (SVM) scores where applied (whereas we utilize SVM and its variations). The results in [2] showed that in general visual features were not successful to estimate the dominance while their fusion with nonverbal audio features usually performed better than audio only. Whereas in [20], audio-visual fusion did not yield any better performance than audio-only features.

A study using head pose to obtain VFOA to find the visual attention for dominance estimation was presented in [19]. In that study [19], VFOA for a person was labeled manually (in contrast to our study such that VFOA is estimated automatically) and also detected automatically using a Bayesian formulation. As nonverbal features the total received visual attention and looking while speaking were used. The results using both manually and automatically extracted cues showed that audio cues were very powerful while visual cues were not effective as much as audio cues.

The most similar study to ours is [31] (only the method presented in Chapter 6) since it aimed to detect the ELs in a meeting environment and used nonverbal video-based features (although combined with nonverbal audio based features as well) which were extracted from VFOA. In that study [31], ELship detection performance were evaluated in terms of variables: leadership, dominance, competence, and liking which is totally different than our evaluation which utilizes human annotations (and also social psychology questionnaires). The result showed that VFOA detection was performed not very sufficiently (42% accuracy) which might be the reason of poor performance (except dominance) of nonverbal visual features compared to audio features.

In this study, we utilize head pose to estimate the eye gaze. The nonverbal visual features are defined using the extracted

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

VFOA. The VFOA extraction is performed automatically and in a supervised way while sufficiently good results are obtained. The most and the least ELs in meeting segments are estimated using SVM and its variants. In this stage, the leadership annotations of human observers (not necessarily full agrement but also based on majority agrement) are used to learn and evaluate the leadership model. The efficiency of proposed nonverbal visual features are evaluated based on human annotations and validated by the social psychology questionnaires.

3. DATASET DEFINITION

The presented dataset consists of 16 meeting sessions such that the longest meeting session lasts 30 minutes while the shortest meeting session lasts 12 minutes (in total 393 minutes). The meeting sessions are composed of the same gender, unacquainted four-person (in total 44 females and 20 males) with average age of 21.6 with 2.24 standard deviation. The participants are seated as given in Figure 1. Videos were recorded using four frontal cameras (with a resolution of 1280x1024 pixels and frame rate of 20 frame per second (fps)) and a standard camera (with a resolution of 1440x1080 pixels and frame rate of 25 fps, used only for data annotation) to capture the whole scene. Audio was recorded with four wireless lapel microphones, each one connected to person's corresponding frontal camera (audio sample rate=16 kHz). The participants performed one "survival task", randomly chosen between two tasks: "winter survival" and "desert survival" [21] which are most common tasks about small group decision making, dominance and leadership. In a typical survival task, participants are presented with a dramatic survival situation in a given geographical layout (e.g., a plane crash in the desert), with some details provided about the general conditions of the context (e.g., time of the day, nearest town distance, etc.). Participants are then given a list of objects that are left after the accident, and their task is to rank each of these items in the order of importance for the survival. A single decision has to be taken by the group and follows a group discussion. In this study, instructions were given verbally, the use of pen and paper was not allowed, and the items to be ordered were 12.

3.1 Questionnaires

The SYstematic method for the Multiple Level Observation of Groups (SYMLOG) [4, 24] is a tool designed to evaluate individual dispositions along three bipolar dimensions: dominance versus submissiveness, acceptance versus non-acceptance of task orientation of established authority, and friendliness versus unfriendliness. The SYMLOG can be used both as a self-assessment instrument and as an instrument for external observation of a group interaction. Before the group task, volunteer participants were asked to complete the SYMLOG questionnaires and it was used to select the designated leaders (DLs) of the task [17]. Subjects with scores of dominance and of task-orientation higher than the median of the sample were selected as DLs. The analysis regarding the DL is beyond the scope of this paper but it is worth to state that a DL may appear as an EL.

After the task, each participant was asked to rate the General Leader Impression Scale (GLIS) [25] questionnaire. The GLIS is an instrument designed to evaluate the leadership attitude that each member displays during a group

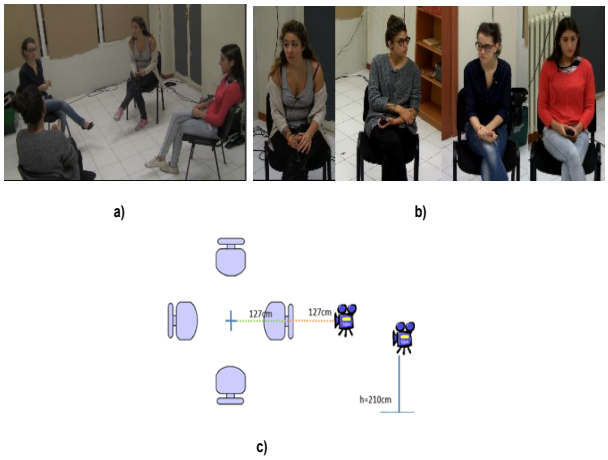

Figure 1: Set up of the dataset. a) Full view of meeting room, b)example of frontal cameras, c) plan of the set up

interaction. It is a 5 item scale which asks participants to rate the other members of the group on their contribution to the group's overall effectiveness on the task. GLIS were calculated for each individual by averaging the ratings given by the other group members.

Additionally, two independent judges observed the meetings and rated each participant of each session using the GLIS (refereed as GLIS-Observers) (InterClass Correlation (ICC)= 0.771; $p<0.001$) and the SYMLOG (refereed as SYMLOGObservers) (dominance ICC= 0.866, task-orientation ICC= 0.569, friendliness ICC= 0.722; $p<0.001$). For SYMLOGObservers only the dominance sub-scale of it, was used since the leadership impression obtained by GLIS-Observers and dominance tend to correlate with each other. The final scores for each participant were calculated as the average between their ratings.

3.2 Data Annotation

16 meeting sessions were divided into small segments (in total 75 segments), each lasting 4, 5 or 6 minutes. All meeting segments were used to analyze the proposed method rather than using the original full meetings. The main reason for segmenting was to be able to have more data for training and testing, in a similar way to [20]. This also resulted in more accurate annotations since people are more precise and more focused on annotation of videos when they were shorter, as mentioned in [1].

Given that, psychology literature found that human observers can identify the ELs [33], in total 50 observers were used to annotate each video segment. Each observer annotated either 12 or 13 video segments (no more than one segment per meeting session). During the annotation, audio was not used in order to overcome any possible problem that might occur due to the level of understanding of the spoken language. Annotators were requested to judge the four participants by ranking them from 1 to 4, where 1 corresponded to the person who exhibited the most leader behavior and 4 corresponded to the person who exhibited the least leader behavior. In this paper, we used the annotations regarding the most and the least EL. The analysis about the annota-

Table 1: Analysis of Leadership Annotations

Emergent Leader	Agreement Type	Average Agreement	Total # of Meetings/ Out of
Most	Full	1	26/75
	Majority	0.73	49/75
Least	Full	1	13/75
	Majority	0.70	62/75

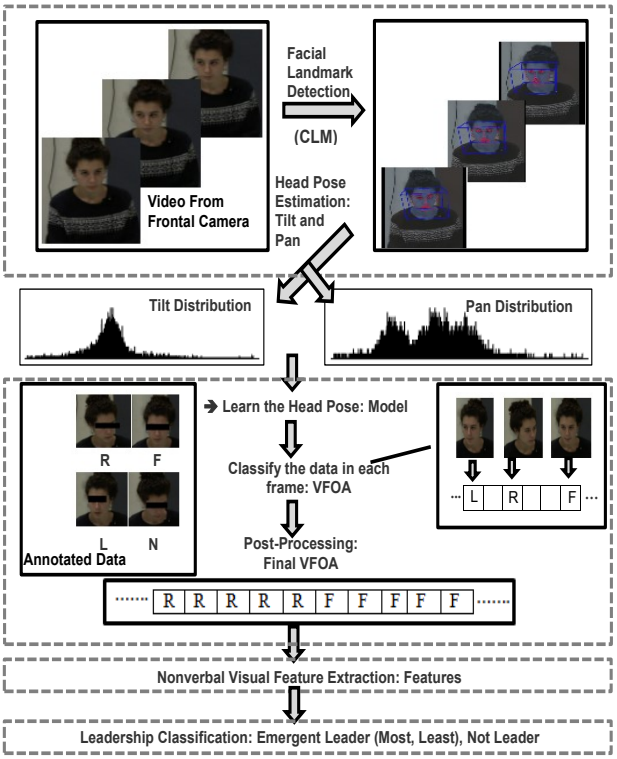

Figure 2: Overview of the Proposed Method

tions is given in Table 1. As seen from Table 1, annotating the least EL was more challenging than annotating the most EL.

4. PROPOSED METHOD

The proposed method (Figure 2) is divided into four parts. First, facial landmark detection and head pose estimation are applied. Then, VFOA is modeled and estimated. Later, nonverbal visual features are extracted. As a result of these steps, the pan, tilt and roll angles of a person for each video frames are obtained. Using the labeled VFOA of a person and a supervised learning algorithm, the entire VFOA of that person is found. Finally, the nonverbal features are extracted from VFOA and used to detect the ELs.

4.1 Facial Landmark Detection and Head Pose Estimation

Facial landmark detection and tracking are based on the Constrained Local Model (CLM) [11]. This method can be briefly summarized as follows: first, a model of faces is built from a training set by using shape (facial landmarks) and texture (patches around landmarks) information; then, the model is fit to a test image through an iterative algorithm, in which, at each iteration, the result of the correlation between the model's patches and the patches (called templates) sampled from the test image feature points is maximized and new feature points are chosen accordingly for the next iteration. When the algorithm converges, the resulting facial landmarks in 2D coordinates are converted to 3D coordinates and used to detect the head pose (pan, tilt, roll) and position in camera space [5].

4.2 Modeling Visual Focus of Attention

A person's VFOA can be defined as a person, object or, more generally, any location the person is looking at [34]. One way of inferring the VFOA is to use the person's eye gaze which is found by detecting and tracking the eyes. Current eye gaze tracking techniques are still constraining [3] and challenging. For instance, they require the person to be close to the camera to track the eyes accurately [16]. In many studies such as [3, 34, 26], it has been shown that the eye gaze can be estimated using the head pose representation.

In this paper, we also use the head pose representation to find the VFOA. The pan and tilt angles are used to define the head pose which is in contrast to studies [34, 8] that utilized only head pan angle while the roll angle is not used (like [3]) since there is no effect of it to head direction. The VFOA of a person contains the other three persons who are on his/her right, left or front (shown as R, L and F, respectively, in Figure 2) and also no-one (shown as N in Figure 2) which refers to the time that the person is not looking to any participants but somewhere else such as ceiling, floor, door, etc. It is important to highlight here that, in this V-FOA definition, all the physical locations different than any other participant are considered as the same class.

In the literature, there are many supervised and unsupervised methods to estimate the VFOA in a meeting environment from head pose representation [3, 34]. In this paper, SVM was used to learn and predict VFOA since it was significantly the best performing method among the compared state of the art methods (see Section 5). Before applying SVM to find the VFOAs, the head pose representations were first interpolated using spline interpolation since there are frame drops in different videos belonging to same meeting and the videos should be synchronized to extract nonverbal features.

To train the SVM classifiers (one for each of 64 frontal videos) and also to evaluate its performance, the VFOA for a total of 25600 randomly selected frames (400 frames for each video which was determined by the confidence level=90% and margin error=4%) were annotated by two annotators. In total 23000 frames (in average 359.4 per video with standard deviation of 46.54) were used for evaluation which were obtained after removing differently labeled V-FOAs. The VFOA annotation results show that we have highly imbalance VFOA classes when the least represented class is no-one which is 16% of the data. The labeled VFOA data were randomly divided into two folds (while having totally different but the same amount of instances from each classes) as training and validation sets and this process was repeated for 100 times to learn the individual SVM models. As SVM model, the radial basis kernel function (RBF) with varying kernel parameter was selected. As stated in [6], SVM tended to be biased towards to well-presented class. To

handle this class imbalance problem, the cost function [13] (SVM-cost), the random under sampling [37] (SVM-RUS) and the SMOTE [10] (SVM-SMOTE) methods were combined with SVM. To evaluate the performance of SVMs, the geometric mean of detection rates (see [6] for definition) were used. For each video, the method (SVM, SVM-cost, SVM-RUS or SVM-SMOTE), performing the highest geometric mean of the detection rates with corresponding parameters was selected to classify the whole unlabeled head pose. This results in VFOA per person for the entire video.

4.3 Nonverbal Visual Feature Extraction

A fixation happens when a participant looks at another participant for a minimum amount of time. In our analysis, fixation was taken as 5 frames and all the VFOAs were smoothed with it as a *post-processing step* to denoise the VFOAs before extracting the features. From the obtained VFOAs for each person the following nonverbal features are extracted.

totWatcher: The total number of frames that a person is being watched by the other persons in the meeting. **totME:** The total number of frames that a person is mutually looking at any other persons in the meeting (also called mutual engagement (ME)). **totWatcherNoME:** The total number of frames that a person is being watched by any other persons in the meeting while there is no ME. **totNoLook:** The total number of frames that are labeled as no-one in the V-FOA vector meaning that a person is not looking at any other persons in the meeting. **lookSomeOne:** The total number of frames that a person looked at other persons in the meeting. **totInitiatorME:** The total number of frames to initiate the MEs with any other persons in the meeting. **stdInitiatorME:** The standard deviation of the total number of frames to initiate the MEs with any other persons in the meeting. **totInterCurrME:** The total number of frames intercurrent between the initiation of ME with any other persons in the meeting. **stdtInterCurrME:** The standard deviation of the total number of frames intercurrent between the initiation of ME with any other persons in the meeting. **totWatchNoME:** The total number of frames that a person is looking at any other persons in the meeting while there is no ME. **maxTwoWatcherWME:** The maximum number of frames that a person is looked at by any other two persons while that person can have a ME with any of two persons. **minTwoWatcherWME:** The minimum number of frames that a person is looked at by any other two persons while that person can have a ME with any of two persons. **maxTwoWatcherNoME:** The maximum number of frames that a person is looked at by any other two persons while that person can have no ME with any of two persons. **minTwoWatcherNoME:** The minimum number of frames that a person is looked at by any other two persons while that person can have no ME with any of two persons. **ratioWatcherLookSOne:** The ratio between the *totWatcher* and *lookSomeOne*.

In total 15 features were extracted. All features (except *ratioWatcherLookSOne*) were divided by the total number of frames in a given meeting since the total number of frames per meeting is variable. The features *totWatcher*, *LookSomeOne* and *ratioWatcherLookSOne* were already used in [31] by combining with nonverbal audio features for EL detection and also in [19] to detect the dominant person in a meeting. To the best of our knowledge the rest of the features were never used in a SSP study, although they have been discussed in social psychology works related to dominance, leadership and nonverbal behavior. In addition to these features, the total number of frames that a person is looked by all other three persons in the meeting with/without a ME can also be extracted. However, for our dataset, we observed that, such features are not useful since there were no such a frame.

The motivation of the extracted features is as follows [9, 15]: how many times and how long *i)* the EL is looked at by each person while there is no ME is a measure of the individual coordination to the leader, *ii)* the EL is looked at by the two members simultaneously when there is no ME is a measure of the group coordination of the leader, and it is expected that higher values of this index reflects the centrality of the leader, in other words, a person is looked at by another two persons simultaneously without ME reflect the group behavior towards an individual person and higher values of this feature could reflect the EL. *iii)* a peer is looked at by the leader without ME reflect's the leader's directiveness and correlate with the perceived efficacy of the leadership at the group level. *iv)* ME is a measure of the reciprocal engagement among the participants, higher values of this feature should reflect better leader-to-peer coordination. *v)* Being initiator of a ME can be seen as a measure of the ability to attract the attention of a person and it is expected that having high values of being initiator reflects the EL's directive activity.

Using the extracted nonverbal visual features, the most and the least EL for each meeting segment were modeled and detected by the methods given in Section 5.2.

5. RESULTS

In this section, we present *i)* the results correspond to different VFOA detection algorithms, *ii)* EL detection results by different algorithms and *iii)* the correlation analysis between each nonverbal feature and the questionnaires.

5.1 Results of VFOA Estimation

Different than SVM and its variations (SVM-cost [13], SVM-RUS [37] and SVM-SMOTE [10]), we applied methods based on OTSU [29], k-means and Gaussian Mixture Model (GMM) [34] to model and to estimate the VFOA. These methods are briefly summarized as follows:

OTSU [29] based method: Pan and tilt angles per a frontal video (in other words per person) were first smoothed assuming that they can vary from -90 to 90 degrees. Then OTSU thresholding was applied to smoothed pan and tilt angles independently. This resulted in four thresholds (two for pan angles and two for tilt angles).

k-means based method: The median and standard deviation of the tilt angles per frontal video were used to define the two thresholds which were obtained as median of tilt angles $\pm$ standard deviation of tilt angles. The pan angles per frontal video were clustered using k-means with three clusters. This resulted in three centers and the two thresholds were found by finding the middle point of the two consecutive cluster centers.

GMM [34] based method: The thresholds from tilt angles per frontal video were obtained as given in k-means based method. Pan angles per frontal video were modeled using GMM with three components (representing left, right and front). The mean and covariance of components were ini-

Table 2: VFOA Estimation

Method / Detection Rate	Right	Left	Front	No-one
OTSU	0.44	0.53	0.55	0.60
k-means	0.75	0.87	0.79	0.10
GMM	0.73	0.77	0.62	0.10
SVM	0.88	0.86	0.67	0.39
SVM-cost	0.85	0.85	0.72	0.52
SVM-RUS	0.83	0.82	0.70	0.56
SVM-SMOTE	0.87	0.86	0.70	0.51

tialized using k-means (having three clusters) where priors were set to uniform.

For all methods, VFOAs (right, left, front and no-one) from head poses per frame were classified as follows: A tilt value (which were obtained using the tilt angles from a frontal video) was classified as no-one if the tilt value was out of the thresholds. For OTSU and k-means, if the tilt value was between the thresholds, then the corresponding pan value was compared with the thresholds obtained using pan angles. If the pan value was smaller than the smallest pan threshold, the VFOA was classified as left; if the pan value was greater than the biggest pan threshold then the VFOA was classified as right; and finally if the pan value was between the pan thresholds then the VFOA was classified as front. For GMM, the maximum class probability was used to estimate the VFOA.

These methods were also combined with some pre-processing steps: 5% outlier removal and smoothing (by moving average filter) which were applied before calculating the thresholds that were obtained from pan and tilt angles (applying the outlier removal and smoothing always improved the results). All the results regarding VFOA estimation are given in Table 2 in terms of detection rate which is defined as follows.

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

where c refers to class.

As seen, SVM and its variations performed better than other methods especially to detect right, left and front. The detection rate of no-one by SVM and its variations was also better than the rest except OTSU which on the other hand performed very poorly to estimate the right, left and front. k-means and GMM were also performed almost as good as SVM and its variations (for detecting right, left, and front) however their no-one detection rate was very low. On the light of those results as mentioned in Section 4.2, SVM and its variations were used to model and estimate the VFOAs.

5.2 Results of Emergent Leader Detection

The variations of SVM (all with RBF with varying kernel parameters) using leave-one-out, leave-one-meeting-out and leave-one-meeting-segment-out approaches and rank-level fusion approach (RLFA) [31, 2] using different feature groups were used to detect the most and the least ELs using the proposed nonverbal visual features.

In Table 3, the best results for SVM (which is selected by the highest score of the geometric mean of the detection rates) and its variations and RLFA with different features were compared when the three classes (the most EL, the least EL and the other persons) were considered. As variations of SVM, SVM-cost [13], SVM using the features after principal component analysis (PCA) was applied (SVM-afterPCA), SVM-cost [13] using the PCA applied features (SVM-afterPCA-cost) and SVM which was applied using the features that were found correlated with the questionnaires (SVM-with-CorrFea, see 5.2.1) were used. For SVM, only the results with leave-one-meeting-out approach is given since all the results were similar to each other. Assuming that the proposed nonverbal features can be correlated with each other which might effect the performance of SVM negatively, PCA was applied to the features as a dimensionality reduction technique. To obtain a useful set of components the smallest number of components that represent 90% of the sum of all eigenvectors was used. This left five features from the defined 15 features. The RLFA was applied using the whole nonverbal features and only with the features correlated with the questionnaires (RLFA-with-CorrFea).

Table 3: Emergent leader (EL) detection performances using nonverbal visual features

Method / Detection Rate	Most EL	Least EL	Rest
SVM	0.71	0.59	0.75
SVM-cost	0.80	0.58	0.70
SVM-afterPCA	0.72	0.63	0.71
SVM-afterPCA-cost	0.79	0.63	0.64
SVM-with-CorrFea	0.67	0.62	0.72
RLFA	0.71	0.71	0.69
RLFA-with-CorrFea	0.72	0.67	0.68

As seen in Table 3, the best performing method for the most EL detection was SVM-cost while its least EL detection rate was the worst. Using PCA improved the detection rate of the least EL. Applying the cost function which penalize the mis-detection of the most and the least ELs more than the rest improved the detection rate of the most EL. The best performing method to detect the least EL was RLFA which in general performed as good as SVM and its variations although it is an unsupervised learning algorithm (similar to the results given in [33, 31, 20]). Overall, the best performing method can be considered as the method which performs well to detect the most EL while not performing poor in detecting the least EL and the rest as well. With such an assumption all methods performed almost the same with ± 0.02 deviation.

Different from the results given here, SVM and its variations were also applied using binary classes as: *i)* the most EL versus the rest and *ii)* the least EL versus the rest. For the detection rate of the most and the least ELs, the results were very similar to the results given in Table 3 while the detection rate of the rest were highly increased (in average 15%) no matter which cross validation approach (leave-one-out, leave-one-meeting-out and leave-one-meeting-segment-out) was applied.

To better investigate the performance of each nonverbal visual features for the most and the least ELs detections, RLFA and SVM-cost were applied using leave-one-meeting-out when the three classes were considered. Additionally, the features (totWatcher, sum of totWatchNoME and totME per frontal video, ratioWatcherLookSOne) used in [31] (shown as Fea-[31]) were also evaluated. The results are given in Table 4.

The results in Table 4 are the best results according to

Table 4: Individual performance of nonverbal visual features for the most and the least emergent leaders

Nonverbal Visual Features	RLFA most	RLFA least	SVM-cost most	SVM-cost least
totWatcher	0.71	0.68	0.74	0.55
totME	0.74	0.68	0.75	0.54
totWatcherNoME	0.68	0.68	0.76	0.54
totNoLook	0.24	0.15	0.38	0.26
lookSomeOne	0.26	0.23	0.38	0.26
totInitiatorME	0.46	0.46	0.50	0.20
stdInitiatorME	0.27	0.22	0.27	0.14
totInterCurrME	0.16	0.29	0.34	0.30
stdtInterCurrME	0.35	0.31	0.36	0.14
totWatchNoME	0.04	0.06	0.75	0.55
maxTwoWatcherWME	0.66	0.67	0.70	0.55
minTwoWatcherWME	0.60	0.60	0.59	0.50
maxTwoWatcherNoME	0.63	0.66	0.67	0.55
minTwoWatcherNoME	0.62	0.50	0.60	0.39
ratioWatcherLookSOne	0.72	0.67	0.72	0.57
Fea-[31]	0.71	0.67	0.52	0.57

geometric mean of detection rates. These results show that the best features to detect the most EL accurately are: totWatcher, totME, totWatcherNoME, maxTwoWatcherWME, maxTwoWatcherNoME, and ratioWatcherLookSOne. The best features to detect the least EL more accurately are totWatcher, totME, totWatcherNoME, maxTwoWatcherWME, maxTwoWatcherNoME, ratioWatcherLookSOne and Fea-[31]. Furthermore, using all features together (see Table 3) performed better for both classes in general. On the other hand, when the performance of the proposed features and the features presented in [31] were compared, it has seen that the most EL detection performance of the proposed features was better no matter which classifier was applied while the least EL detection rates were similar.

5.2.1 *Correlation Analysis*

In Table 5, the correlation between variables derived from questionnaires and visual features are given when the meeting videos were evaluated as whole, rather than segmented, as defined in Section 3.2. As seen from Table 5, except totNoLook, lookSomeOne, stdInitiatorME all other nonverbal features found correlated (eight of them had high correlation, two of them had medium correlation and two of them had low correlation) with the results of SYMLOG-Observers. Similarly, except totNoLook, lookSomeOne, stdInitiatorME and totInterCurrME all other nonverbal features were correlated (seven of them had high correlation, three of them had medium correlation and one of them had low correlation) with the results of GLIS-Observer.

Table 5: Correlation Coefficient Values Between Questionnaires and Nonverbal Visual Features

Nonverbal Visual Features	SYMLOG-Observers	GLIS-Observers
totWatcher	0.69	0.68
totME	0.61	0.59
totWatcherNoME	0.67	0.66
totNoLook	0.06	-0.08
lookSomeOne	-0.06	0.08
totInitiatorME	0.31	0.42
stdInitiatorME	0.005	0.08
totInterCurrME	-0.20	-0,06
stdtInterCurrME	0.23	-0.14
totWatchNoME	-0.61	-0.49
maxTwoWatcherWME	0.65	0.60
minTwoWatcherWME	0.51	0.52
maxTwoWatcherNoME	0.52	0.50
minTwoWatcherNoME	0.44	0.48
ratioWatcherLookSOne	0.65	0.59

6. CONCLUSIONS

In this work, we presented novel nonverbal visual features which are extracted from VFOA to detect the ELs in a meeting environment. Different than many ELship studies, we only used video cues although it was shown that audio cues were generally more effective. The proposed nonverbal features performed well for detection of the most and the least ELs (70% of detection rate in average) when the majority of the defined nonverbal features were highly correlated with the results of the social psychology questionnaires. The human annotations using the video segments showed very high overlap (94% overlap with SYMLOG-Observers for the most and the least leaders, and 88% overlap with GLIS-Observers for the most and the least leaders when the highest/lowest values of the questionnaires were used for EL inference) with the results of questionnaires which were filled by observers using the whole videos. In the 58 out of 75 video segments, the most EL annotated by the 50 human observers was also the DL. Similarly, in 12 out of 16 whole videos, the most EL inferred by GLIS-Observers was also the DLs. The applied supervised and unsupervised methods to detect the most and the least ELs performed well, which can be a result of the accurate detection of VFOAs (72% detection rate in average) and the effectiveness of the used features.

As future work, novel nonverbal audio features will be defined and fused with the nonverbal visual features and the accuracy of the single-modality and the multi-modalities will be compared. Moreover, to determine the ELship, the interactions between persons in a meeting environment will be modeled as sequences of events using audio and video cues.

7. REFERENCES

[1] N. Ambady, F. Bernieri, and J. Richeson. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, 32:201–257, 2000.

[2] O. Aran and D. Gatica-Perez. Fusing audio-visual nonverbal cues to detect dominant people in small group conversations. In *ICPR*, pages 3687–3690, 2010.

[3] S. O. Ba and J.-M. Odobez. Recognizing people's focus of attention from head poses: a study. *IDIAP Research Report 06-42*, pages 1–27, 2006.

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

[5] 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*, pages –, 2013.

[6] C. Beyan and R. Fisher. Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, 48(5):1653–1672, 2015.

[7] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic. The mahnob mimicry database - a database of naturalistic human interactions. *Pattern Recognition Letters*, 66:52–61, 2015.

[8] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal. The AMI meeting corpus: A pre-announcement. In *MLMI*, pages 28–39, June 2005.

[9] D. R. Carney, J. A. Hall, and L. S. LeBeau. Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29(2):105–122, 2005.

[10] V. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[11] D. Cristinacce and T.F.Cootes. Feature detection and tracking with constrained local models. In *BMVC*, pages 929–938, 2006.

[12] S. Feese, A. Muaremi, B. Arnrich, G. Troster, B. Meyer, and K. Jonas. Discriminating individually considerate and authoritarian leaders by speech activity cues. In *IEEE PASSAT, and IEEE SocialCom*, pages 1460–1465, 2011.

[13] G. Fumera and F. Roli. Cost-sensitive learning in support vector machines. In *the Workshop Mach. Learn. Meth. Appl.*, pages –, 2002.

[14] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest level in meetings. In *IEEE ICASSP*, pages 489–492, 2005.

[15] J. A. Hall, L. S. LeBeau, and E. J. Coats. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, 131(6):898–924, 2005.

[16] D. Hansen and Q. Ji. In the eye of the beholder: a survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(3):478–500, 2010.

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

[18] 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*, 19(4):847–860, May 2011.

[19] 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 *ICMI*, pages 233–236, 2008.

[20] 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., Sp. Issue on Multimodal Processing for Speech-based Interactions*, 17(3):501–513, 2009.

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

[22] N. Jovanovic, R. op den Akker, and A. Nijholt. Addressee identification in face-to-face meetings. In *EACL*, pages 169–176, 2006.

[23] M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal Communication in Human Interaction.* 8th Edition, Wadsworth, Cengage Learning, Boston, 2013.

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

[25] R. Lord, R. Foti, and C. D. Vader. A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational behavior and human performance*, 34(3):343–378, 1984.

[26] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Here's looking at you, kid. detecting people looking at each other in videos. In *BMVC*, pages –, 2011.

[27] 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.*, 27(3):305–317, March 2005.

[28] T. Meservy, M. Jensen, J. Kruse, J. Burgoon, J. Nunamaker, D. Twitchell, G. Tsechpenakis, and D. Metaxas. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20(5):36–43, 2005.

[29] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Syst., Man, Cybern., Syst.*, 9(1):62–66, 1979.

[30] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In *ICMI*, pages 255–262, 2007.

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

[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*, 7(1–2):39–53, August 2012.

[33] 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. On Multimedia*, 14(3):816–832, 2012.

[34] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. on Neural Networks*, 13(4):928–938, 2002.

[35] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi. Putting the pieces together: multimodal analysis of social attention in meetings. In *ACM Multimedia*, pages 25–29, October 2010.

[36] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *CVIU*, 143:11–24, 2016.

[37] B. Yap, K. Rani, H. Rahman, S. Fong, Z. Khairudin, and N. Abdullah. An application of oversampling, undersampling, bagging, and boosting in handing imbalanced datasets. *In DaEng, Lecture Notes in Electrical Engineering*, 285:13–22, 2014.