Identification of Emergent Leaders in a Meeting Scenario Using Multiple Kernel Learning

Cigdem Beyan[1] Francesca Capozzi[2] Cristina Becchio[3,4] Vittorio Murino[1,5]
[1]Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy
[2]Department of Psychology, McGill University, Montreal, QC, Canada
[3]Department of Psychology, University of Turin, Italy
[4]Robotics, Brain and Cognitive Sciences, Istituto Italiano di Tecnologia, Genova, Italy
[5]Department of Computer Science, University of Verona, Verona, Italy
(cigdem.beyan,cristina.becchio,vittorio.murino)@iit.it, francesca.capozzi@mcgill.ca

ABSTRACT

In this paper, an effective framework for detection of emergent leaders in small group is presented. In this scope, the combination of different types of nonverbal visual features; the visual focus of attention, head activity and body activity based features are utilized. Using them together ensued significant results. For the first time, multiple kernel learning (MKL) was applied for the identification of the most and the least emergent leaders. Taking the advantage of MKL's capability to use different kernels which corresponds to different feature subsets having different notions of similarity, significantly improved results compared to the state of the art methods were obtained. Additionally, high correlations between the majority of the features and the social psychology questionnaires which are designed to estimate the leadership or dominance were demonstrated.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Emergent leadership; head pose; head activity; body activity; multiple kernel learning; social signal processing

1. INTRODUCTION

Social interaction is the main facet of human life and also the fundamental research area for social psychology. Even though psychologists have been working on social interactions for a very long time, the automatic analysis of them is a relatively new problem.

Social interactions are based on verbal (the words spoken) and nonverbal (such as eye gaze, head-body activities, body gestures, facial expressions, speaking style, speaking time, interruptions, turn, etc.) communications. Verbal communication is a key in social interactions but it is also known that a high amount of information is conveyed nonverbally [18]. Besides, nonverbal communication allows to estimate the human personality and relationships [18].

Social signal processing (SSP) is the field which aims to analyze human interactions in an automatic way using the recent advances in machine analysis (e.g. speech processing, computer vision, machine learning). One topic among many others in this area covers just social interactions in a small group [25] such as a meeting. As examples, detecting group interest level during meetings [10], modeling dominance in small group conversations [2], identifying emergent leaders (EL) in a meeting scenario [24, 6], and analysis of social focus of attention [28] can be given.

In this paper, we are investigating the small groups in terms of emergent leadership. An emergent leader (EL) is the person who appears as the leader during a social interaction. His/her leading power is related to his/her dominance, influence, leadership and control rather than his/her role in the organizational hierarchy [27]. Automatically identifying ELs in a small group first investigated in [24] and in the publications (such as [26, 27, 25]) related to that work. Recently, in [6], the most and the least ELs in a meeting environment were detected using visual nonverbal features only.

To automatically identify the ELs, different supervised (Support Vector Machine (SVM) [6], ranking the scores of SVM [27, 24], collective classification [27, 24]) and unsupervised learning (rule-based approach [27, 25, 24, 26], rank-level fusion approach (called as RLFA in this paper) [26, 27, 24]) methods have been used. To the best of our knowledge, none of the existing works applied multiple kernel learning (MKL) for detection of ELs. However, given that different modalities and different types of features are being used, MKL can perform well thanks to its ability to use different kernels corresponding to different feature subsets having different notions of similarity.

In this work, we combine different types of visual nonverbal features to detect the most and the least ELs in a meeting environment. The performance of MKL is compared with the state of the art methods. The main contribution of this work is utilizing MKL (for the first time) to detect the ELs which demonstrated significantly improved results. Additionally, for the first time, VFOA, head activity and body activity based features are used together for emergent leadership. Unlike the majority of the works in emergent

ASSP4MI'16, November 16 2016, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3005467.3005469

leadership literature, only visual features are utilized which showed good results. This is also promising especially to analyze the social interactions in the absence of audio sensors [6, 14]. The presented results are also considerable regarding the existing studies which showed better EL estimation using audio features compared to video features or audio-video together. Furthermore, it is likely that their fusion with audio features might produce much better results when the visual features are extracted using the methods given in this work.

The rest of this paper is organized as follows. The previous studies about leadership are discussed in Section 2. In Section 3, the data collection including questionnaires used and the annotation of the data are given. In Section 4, the nonverbal features used are described and the methods used to extract them are presented. The experimental analysis including EL detection results and the correlation analysis of each feature with the questionnaires are reported in Section 6. Finally, we conclude the paper with discussions and future work in Section 7.

2. RELATED WORK

In this section, emergent leadership studies in SSP were reviewed. The nonverbal features they used and the learning methods that they applied were discussed. For the interested readers, a detailed survey on emergent leadership in social psychology and social computing can be found in [24].

In [26], only speech activity was used to identify emergent leader. As nonverbal features; speaking length, speaking turn duration and successful interruptions were used. To predict the EL, a rule-based estimator (which is based on the hypothesis that the EL is the one who has the highest value of a nonverbal feature) and RLFA (which tests if any combination of features are performing better than a single feature) were used. The evaluation of the nonverbal features were realized using the variables perceived leadership (PLead), perceived dominance (PDom), perceived competence (PComp) and dominance ranking (RDom). The results showed that there is a significant relationship between PDom and emergent leadership and the EL is the person who talks the most, has more turns and interruptions. Moreover, the audio nonverbal features were successful to identify the EL even they used individually (60-80% accuracy).

The fusion of speaking activity based features (speaking length, speaking turns and speaking interruptions) with features based on VFOA (attention received, given attention, attention quotient and attention center) was investigated in [25]. In that paper [25], looking while speaking, looking while listening, being looked at while speaking, center of attention while speaking and visual dominance ratio (the ratio of looking while speaking and looking while listening without speaking [14]) were extracted to identify the ELs. Similar to [26], a rule based estimator was used. The evaluation of features were carried out in terms of PLead, PDom, and RDom. The results showed that speaking activity based features are not only better than visual attention based features but also better than the fusion of these multi modalities to detect the ELs. The main reason of poor performance of nonverbal visual features can be the insufficient performance of the method used to extract VFOA automatically (42% frame level accuracy on a subsample of the data). In [24] (only the method presented in Chapter 6), the emergent leadership detection performance of the same visual attention features were explored more deeply in terms of variables PLead, PDom, PComp, perceived liking (PLike) and RDom. According to their analysis, the amount of attention received from participants was the most informative feature while the attention center was the following best feature for emergent leadership. Additionally, for attention received a strong correlation with PLead, PDom and RDom were detected. On the other hand, a negative correlation for attention received and attention center with PLike were found.

In [27], the emergent leadership were addressed using audio and video based nonverbal features individually and together. As nonverbal audio features; speaking turn features (such as total speaking length and total speaking turns) and prosodic features (such as energy and pitch variation) were extracted. As visual nonverbal features; head activity based features (see Section 4.2), body activity based features (see Section 4.3), and motion template based features (mean, median, quantile and entropy of body activity which are extracted from motion energy image and motion history image that summarize the spatial temporal activity of a person) were used. Different than the presented feature extraction in Section 4.2, in [27] head tracking was performed using Particle filter. Unlike our study and [25], VFOA based features were not used. But, the motion template based features which can be seen redundant given that features based on head and body activities were already being used were utilized. As learning and estimation methods; rule based estimation, RLFA, ranking the scores of SVM (with linear kernel), collective classification [21] were compared. The results showed that head and body activity based features performs better than other visual features for PLead, PDom and RDom. For audio features only the RLFA was the best performing method followed by collective classification methods. Moreover, speaking turn and the energy features were the best features in terms of PLead, PDom and RDom accuracies. For audio and visual features together, RLFA performed the best while collective classification methods were as good as it for PLead, PDom and RDom. Speaking turn based features, body activity based features and energy together were the best performing feature combination to infer the ELs.

Unlike studies [26, 25, 27] which utilized audio features or audio with visual features, [6] presented detection of the most and the least ELs in a meeting scenario using visual nonverbal features only. In that study [6], head pose was used to approximate the gaze. Then, VFOA was estimated from head pose and novel VFOA features (see Section 4.1) were proposed. The estimation of VFOA was performed by methods based on SVM (72% frame level accuracy on a subsample of the data). The best performing method when all VFOA based features were used was SVM-cost [9] for the most EL detection and the best performing method to detect the least EL was RLFA.

In this paper, the VFOA, head and body activity based features are combined together to identify the most and the least ELs in a meeting scenario. The combination of different feature types presents very encouraging performance. As the main contribution, for the first time, the MKL was applied which demonstrated improved results compared to the best performing methods of emergent leadership: RFLA and SVM.

3. DATA COLLECTION

The leadership dataset used in this paper is the same dataset presented in [6]. It is in overall 393 minutes having 30 minutes meeting as the longest session and 12 minutes meeting as the shortest session. There are 16 meeting sessions in total while each meeting sessions are composed of the same gender, unacquainted four-participant (in total 44 females and 20 males with average age of 21.6 with 2.24 standard deviation). In total five cameras; four frontal (with a resolution of 1280x1024 pixels and frame rate of 20 frame per second) to capture each participant individually and a standard camera to record the whole scene (with a resolution of 1440x1080 pixels and frame rate of 25 frame per second, used only for data annotation) were used. Four wireless lapel microphones were utilized for audio acquisition (audio sample rate=16 kHz). The usage of audio is out of the scope of this paper and will be investigated as future work. The participants performed either "winter survival" or "desert survival" [16] tasks as these are the most common tasks in small group decision making, dominance and leadership. For more details see [6].

3.1 Questionnaires

In total two different questionnaires namely *i)* the SYstematic method for the Multiple Level Observation of Groups (SYMLOG) [4, 19] and *ii)* the General Leader Impression Scale (GLIS) [20] were used for evaluation. The SYMLOG is a tool to evaluate individuals in terms of dominance versus submissiveness, acceptance versus non-acceptance of task orientation of established authority, and friendliness versus unfriendliness. On the other hand, GLIS is an instrument used to evaluate the leadership attitude that each participant displays during a group interaction.

Both tools can be used as a self-assessment instrument and also as an instrument for external observation of a group interaction. The interested readers can refer to [6] to get information about how and why SYMLOG and GLIS were applied as a self-assessment instrument. In this paper, the results obtained from external judges were used to evaluate the extracted nonverbal visual features in terms of emergent leadership. In detail, two independent judges were used to observe each meeting and rate each participant using SMYLOG (called as SYMLOG-Observers in this paper) and GLIS (called GLIS-Observers in this paper). The results showed that the dominance inter-class correlation (ICC), task orientation ICC and friendliness ICC were 0.866, 0.569 and 0.722, respectively while $p<0.001$ for SYMLOG-Observers. For leadership impression, in our analysis we only used the dominance sub-scale of SYMLOG. For GLIS-Observers, ICC was 0.771 when $p<0.001$ for the leadership attitude that each member displays during a group interaction. Additionally, it was observed that the leadership impression obtained by GLIS-Observers and the dominance from SYMLOG-Observers tend to correlate with each other. The final scores of each questionnaire for each participant were calculated as the average between the ratings of two observers.

3.2 Data Annotation

16 meeting sessions were divided into small segments, each lasting 5 minutes in average. By this segmentation, in total, 75 meeting segments were obtained. For the analysis presented in Section 6.1, these 75 meeting segments were used rather than using the original full meetings. The reasons of such a segmentation were *i)* to obtain more training and testing data similar to the approach in [15] and *ii)* to have more accurate ground truth annotations since people are more precise and stay more focused on annotation of shorter videos as mentioned in [1].

For annotation of meeting segments, in total, 50 observers were participated. Each observers annotated either 12 or 13 meeting segments in total, while no more than one segment which belongs to the same meeting session was annotated by the same observer. Each meeting segment was annotated by 8 annotators in average. Here, it is important to highlight that psychology literature has already shown that human observers are able to identify the ELs in a meeting scenario [27]. During annotations, audio was not used to cope with any possible problem that might occur due to the level of understanding the spoken language (similar to [17]) which also allowed us to utilize international observers.

Observers ranked the emergent leadership behavior that each participant exhibited in a meeting session. In this paper, we used the annotations regarding the most and the least EL. The other rankings considered as the same class (called the rest in Section 6.1). The analysis of leadership annotations showed that annotating the least EL was more challenging than annotating the most EL. In detail; for the most EL annotation, in the 26 out of 75 video segments, there were fully agreement and in the 49 out of 75 video segments, there were 73% agreement. On the other hand, for the least leader detection annotation, in the 13 out of 75 video segments there were a 100% agreement while in the 62 out of 75 video segments there were 70% agreement. For each meeting segment Krippendorff's α coefficient (in total 75) was also calculated using annotations: the most, the least Els and the rest. The average of Krippendorff's α was found as 0.51 (reliability exist) with 0.27 standard deviation while 7 segments have α smaller than 0.10 (low reliability) and 6 segments have α which is equal to 1.00 (perfect reliability).

4. NONVERBAL FEATURE EXTRACTION

In this section, the description of the extracted nonverbal visual features and the methods used to obtain them are presented. These nonverbal visual features include *i)* the visual focus of attention (VFOA) based features which were extracted using the estimation of the head pose, *ii)* the head activity based features which were extracted using face detection and optical flow, and *iii)* the body activity based features that were obtained using image differencing. The feature extraction process for each type of features are given in Figures 1, 2 and 3.

4.1 Visual Focus of Attention Based Features

To extract VFOA based features, the approach given in [6] was utilized. This method includes facial landmark detection, head pose estimation, modeling VFOA in a supervised way, and estimating the whole VFOA to extract nonverbal features.

By using the Constrained Local Model (CLM) [8], the facial landmarks in 2D coordinates were converted to 3D coordinates which were used to detect the head pose (pan, tilt and roll). Later, the head pose representation (pan and tilt only) was used to find the VFOA. The VFOA of a participant is composed of four possibilities: left if the participant

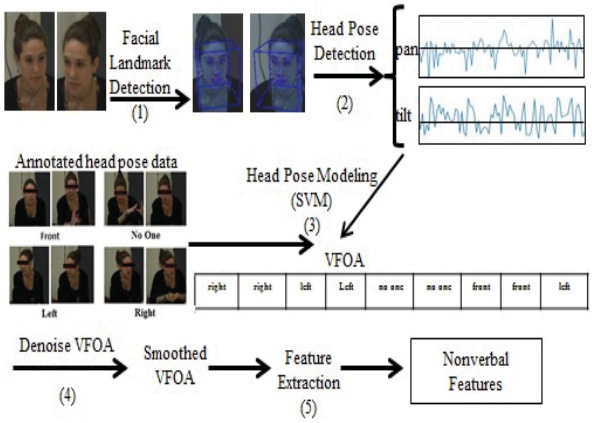

Figure 1: Extraction of VFOA based features

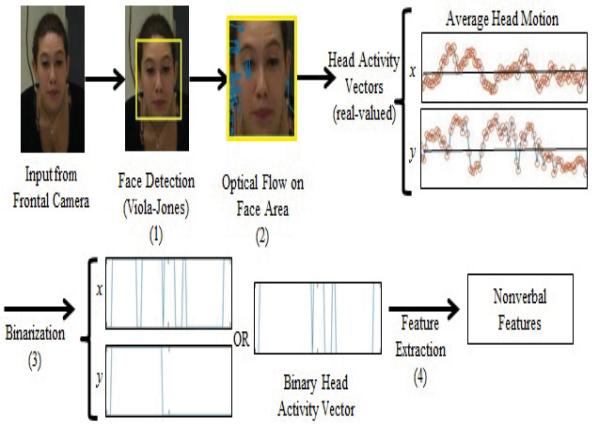

Figure 2: Extraction of head activity based features

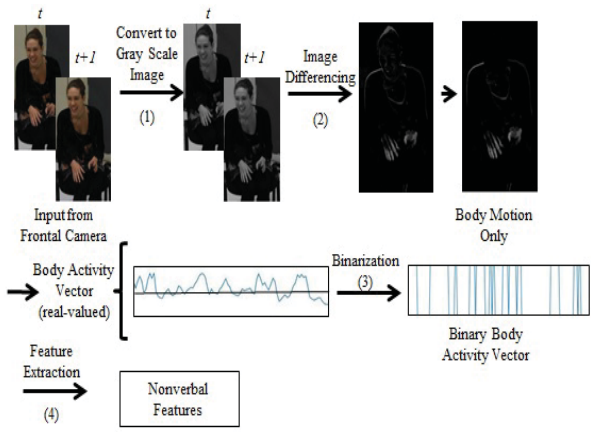

Figure 3: Extraction of body activity based features

is looking at the participant on his/her left, right if the participant is looking at the participant on his/her right, front if the participant is looking at the participant on his/her front, no-one if the participant is not looking at any other participant but somewhere else.

For modeling and estimating a participant's VFOA for the entire video, the cost function [9] (SVM-cost), the random under sampling [31] (SVM-RUS), and the SMOTE [7] (SVM-SMOTE) methods were combined with SVM (see [6] for more information). As SVM model, the radial basis kernel function (RBF) with varying kernel parameters were used. After a participant's VFOA was obtained it was smoothed (the span used for the moving average was taken as 5) to denoise. Finally, the following nonverbal features (referred as VFOAFea in Section 6) were extracted as presented in [6]:

$totWatcher_i$**:** The total time that participant i is being watched by the other participants in the meeting.

$totME_i$**:** The total time that participant i is mutually looking at any other participants in the meeting (also called mutual engagement (ME)).

$totWatcherNoME_i$**:** The total time that participant i is being watched by any other participants in the meeting while there is no ME.

$totNoLook_i$**:** The total time that are labeled as no-one in the VFOA vector meaning that participant i is not looking at any other participants in the meeting.

$lookSomeOne_i$**:** The total time that participant i looked at other participants in the meeting.

$totInitiatorME_i$**:** The total time that participant i initiate the MEs with any other participants in the meeting.

$stdInitiatorME_i$**:** The standard deviation of the total time that participant i initiate the MEs with any other participants in the meeting.

$totInterCurrME_i$**:** For participant i the total time intercurrent between the initiation of ME with any other participants in the meeting.

$stdtInterCurrME_i$**:** For participant i the standard deviation of the total time intercurrent between the initiation of ME with any other participants in the meeting.

$totWatchNoME_i$**:** The total time that participant i is looking at any other participants in the meeting while there is no ME.

$maxTwoWatcherWME_i$**:** The maximum time that participant i is looked at by any other two participants while participant i can have a ME with any of two persons.

$minTwoWatcherWME_i$**:** The minimum time that participant i is looked at by any other two participants while participant i can have a ME with any of two participants.

$maxTwoWatcherNoME_i$**:** The maximum time that participant i is looked at by any other two participants while participant i can have no ME with any of two participants.

$minTwoWatcherNoME_i$**:** The minimum time that participant i is looked at by any other two participants while participant i can have no ME with any of two participants.

$ratioWatcherLookSOne_i$**:** The ratio between the $totWatcher_i$ and $lookSomeOne_i$.

In total 15 features were extracted. All features (except $ratioWatcherLookSOne$) were divided by the length of the corresponding meeting since the lengths of the meetings are variable.

4.2 Head Activity Based Features

The extraction of head activity based features were adapted from [27] which were used by many other works such as [3] and [22] to identify different types of social interactions. Different than the method presented in [27] to detect and track the faces, in this study, we used the most well known face detection algorithm called Viola-Jones [23] which is based on Haar-like features and AdaBoost. A trained face detector was used to detect the face of each participant. Basically, this detector detects the faces and give rectangle bounding boxes which tightly surrounds the detected faces. This method was evaluated using 25600 randomly selected frames (400 frames for each frontal video which was determined by the confidence level=90% and margin error=4%) and performed 90% accuracy with standard deviation of 0.12.

After the face area was detected, the optical flow vectors (using Lucas-Kanade optical flow algorithm [5]) of two consecutive frames within the face area were found. The optical flow vectors were used to obtain the average head motion in x and y coordinates. This result in real-valued vectors representing a participant's head activity in 2 dimensions for a given meeting. These vectors were binarized using a threshold to distinguish the significant head activities from less significant ones. The thresholds (one for each dimension) were defined as the sum of the mean and the standard deviation of head motion per dimensions. This results in two binary vectors one for each dimension where the head activity values greater than the threshold represents significant activity for a given dimension and the values smaller than or equal to the threshold represents an insignificant head activity (small movements, noise, etc.). The obtained binary vectors were fused with an OR operation to have a final binary head activity vector.

Instead of using the optical flow vectors, using the absolute displacement of center of face bounding boxes in consecutive frames can be an alternative way to infer the head motion. Our analysis with this approach showed that using absolute displacement is not significantly worse than using optical flow vector to detect the most and the least ELs. But since approach with optical flow vectors performed better in general, for the analysis in Section 6, this approach was used.

Using the obtained real-valued head activity vectors and the binary head activity vector for each participant, the following features (referred as HeadActFea in Section 6) were extracted:

THL_i**:** The total time that the head of participant i is moving.

THT_i**:** The number of head activity turns for participant i where each turn represents a continuous head activity.

AHT_i**:** The average head activity turn duration for participant i.

$stdHx_i$ and $stdHy_i$**:** The standard deviation of head activity for participant i in x and y dimensions, respectively.

In total 5 features were extracted using the head activity. All features except *stdHx* and *stdHy* were calculate using the binary vector that was obtained whereas for *stdH* the real valued vector was used.

4.3 Body Activity Based Features

The extraction of body activity based features were applied as given in [27]. Since in all of the meeting sessions the background is stationary, image differencing was useful to detect the moving pixels which suppose to be belong to the participant in the frontal video. Here, it is possible to use different foreground detection algorithms but our inference (after testing different algorithms) is that image differencing is more practical as it is less sensitive to noise besides being the simplest foreground detection method. All the moving pixels except the detected face area (obtained as given in Section 4.2) was considered as a part of the body.

Before finding the difference image between consecutive frames, each frame were converted to a gray-scale image. The difference image was found in terms of moving and not moving pixels using a threshold (taken as 30). Hence, if the difference between the gray-scale values of two pixels belong to the consecutive frames were greater than the threshold used, that pixel was labeled as a moving pixel, otherwise it was labeled as a not moving pixel. After the moving pixels were found, the total number of moving pixels in each frame were normalized by the size of the frame. As a result of these steps, a real-valued vector was obtained. This vector was binarized using another threshold (taken as %5) to distinguish the significant body activities from less significant ones.

Using the obtained real-valued vector and the binary vector for each participant, the following features (referred as BodyActFea in Section 6) were extracted:

TBL_i**:** The total time that the body of participant i is moving.

TBT_i**:** The number of body activity turns for participant i where each turn represents a continuous body activity.

ABT_i**:** The average body activity turn duration for participant i.

$stdB_i$**:** The standard deviation of body activity for participant i.

In total 4 features were extracted using the body activity. All features except *stdB* were calculated using the binary vector that was obtained while for *stdB* the real valued vector was used.

5. MULTIPLE KERNEL LEARNING

Multiple Kernel Learning (MKL) methods use a set of kernels and learn the optimal combinations of them in either linear or non-linear way. MKL methods in general preferred due to *i)* their ability to find the optimal kernel combination from a large set of kernels instead of trying which kernel works the best, and *ii)* to utilize different kernels which can correspond to different feature subsets coming from multiple sources which probably have different notions of similarity [12]. There have been extensive work on MKL in the literature; for a comprehensive survey on different MKL methods and their comparisons, interested readers can refer to [12].

The simplest way to combine different kernels is to use an un-weighted sum of kernel function which gives equal preferences to all kernels. A better strategy is to learn a weighted sum. Arbitrary kernel weights (linear combination), non-negative kernel weights (conic combination) and weights on a simplex (convex combination) are possible kernel combinations. Linear combinations of weights can be restrictive whereas a nonlinear combination can be better.

In this study, we utilized Localized Multiple Kernel Learning (LMKL) [11, 12] which utilize nonlinear combinations of kernel weights. LMKL is based on assigning different kernel weights to different regions of the feature space. Briefly, this method contains two components such that their optimizations are performed jointly with a two-step procedure.

The first component is based on a gating model to select the locally optimal kernel function by assigning weights to kernel for a subset of data. The second component presents a locally combined kernel matrix for learning. Using two-step procedure, *i)* an optimization is performed using a fixed gating model, and *ii)* the gating model is updated using the gradients calculated by the current solution. The main advantage of this method is that it allows to use the same type of kernel (e.g. linear, polynomial, Gaussian kernels) for different subset of data using a nonlinear gating model which potentially results in better decision boundaries. For more detail see [11].

LMKL can be combined with any kernel based learning algorithm. In this work, it is combined with SVM. Different number of linear kernels was used (from 2 to 5) with a gating model function either sigmoid or softmax. SVM was applied with kernel parameter $C = 2^{i}$, $i = -1, 1, 3...31$ while C is a trade-off parameter between model simplicity and classification error.

6. RESULTS

In this section, we present *i)* EL detection results by different algorithms using different set of nonverbal visual features and their combination, *ii)* the correlation analysis that was performed between each nonverbal feature and the questionnaires that were described in Section 3.1.

6.1 Results of Emergent Leader Detection

The performance of LMKL was compared with; *i)* single-kernel SVM, *ii)* RFLA and *iii)* another popular multiple kernel learning method called Generalized Multiple Kernel Learning (GMKL) [30, 29]. All the methods were applied for binary classification of *a)* the most EL versus the rest (the least EL and the other persons were modeled as the same class), *b)* the least EL versus the rest (the most EL and the other persons were modeled as the same class) and *c)* multi-class classification with 3 classes (the most EL, the least EL and the rest); using one-versus-one binary classifications as suggested in [13]. During these analysis cross validation approaches: leave-one-meeting-out and leave-one-meeting-segment-out were applied.

For MKL methods, as the kernel based learning method SVM was used. For GMKL, sum and product of RBF kernel subject to l_1 and l_2 regularizations were tested while the number of kernels was the same with the number of features (in our case 24). For LMKL, as gating model sigmoid and softmax functions were used with linear kernels in total 2 to 5. For single-kernel SVM, RBF and linear kernel were used. As kernel parameters C was taken as 2^{i}, $i = -1, 1, 3...31$ and RBF γ was used as 2^{j}, $j = -11, -9, -7...11$. C parameter was taken the same for all methods. RFLA which was in overall the best performing method for emergent leadership identification so far, was applied as defined in [24, 26, 27, 25].

In Tables 1, 2 and 3, the best results of each method when leave-one-meeting-out cross validation was applied were given. SVM and RFLA were applied using each feature group individually (shown as VFOAFea, HeadActFea and BodyActFea) and when they were concatenated (shown as All). GMKL and LMKL were applied using all features together. The multi-class classification and the binary class classifications such that the most EL versus rest and the least EL versus rest were given in those tables, respectively.

Table 1: The best results of each method: multi-class classification (the best results are emphasized in bold-face).

Method / Detection Rate	Most EL	Least EL	Rest
VFOAFea-SVM	0.71	0.59	0.75
HeadActFea-SVM	0.39	0.46	0.60
BodyActFea-SVM	0.57	0.64	0.27
All-SVM	0.74	0.72	0.76
VFOAFea-RFLA	0.71	0.71	0.69
HeadActFea-RFLA	0.29	0.35	0.44
BodyActFea-RFLA	0.61	0.57	0.55
All-RFLA	0.64	0.55	0.59
All-GMKL	0.67	0.57	**0.98**
All-LMKL	**0.79**	**0.75**	0.91

Table 2: The best results of each method: the most EL versus the rest classification (the best results are emphasized in bold-face).

Method / Detection Rate	Most EL	Rest
VFOAFea-SVM	0.71	**0.96**
HeadActFea-SVM	0.60	0.57
BodyActFea-SVM	0.60	0.68
All-SVM	0.75	0.93
VFOAFea-RFLA	0.71	0.90
HeadActFea-RFLA	0.29	0.74
BodyActFea-RFLA	0.61	0.86
All-RFLA	0.64	0.88
All-GMKL	0.62	0.91
All-LMKL	**0.83**	0.91

The aim of the analysis presented here, is to understand the improvement of LMKL (if exist) compared to single-kernel SVM, RFLA and GMKL when all features used. Additionally, the contribution of different feature groups (VFOAFea, HeadActivityFea, and BodyActivityFea) were investigated by using single-kernel SVM and RFLA.

The performances of each method are expressed in terms of the detection rate which is defined as follows.

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

where c refers to class which is the most EL, the least EL and the rest in our case. The best results presented in Tables correspond to the highest score that was obtained from the geometric mean of the detection rates of each class.

The results showed that LMKL is significantly better than using single-kernel SVM, GMKL and RFLA for detecting most and the least EL (for the analysis in Tables 1 and 2). Paired t-test analyses between LMKL and SVM, RFLA and GMKL using the geometric mean results when multi-class classification was applied resulted in p-values 0.04, 0.01, and 0.04 respectively that show significance when the significance level is taken as 0.05. For detection of the least EL (in Table 3), the methods GMKL, RFLA when VFOAFea was used and LMKL performed equally well. GMKL performed worse than SVM especially to detect the most EL (Tables 1 and 2).

The performances of LMKL were generally the same when different gating models (softmax or sigmoid) were used.

In overall, VFOAFea-SVM performed better than Head-

Table 3: The best results of each method: the least EL versus the rest classification (the best results are emphasized in bold-face).

Method / Detection Rate	Least EL	Rest
VFOAFea-SVM	0.67	0.76
HeadActFea-SVM	0.52	0.76
BodyActFea-SVM	0.42	0.83
All-SVM	0.64	**0.90**
VFOAFea-RFLA	**0.71**	0.89
HeadActFea-RFLA	0.35	0.76
BodyActFea-RFLA	0.57	0.83
All-RFLA	0.55	0.84
All-GMKL	**0.71**	0.15
All-LMKL	**0.71**	0.89

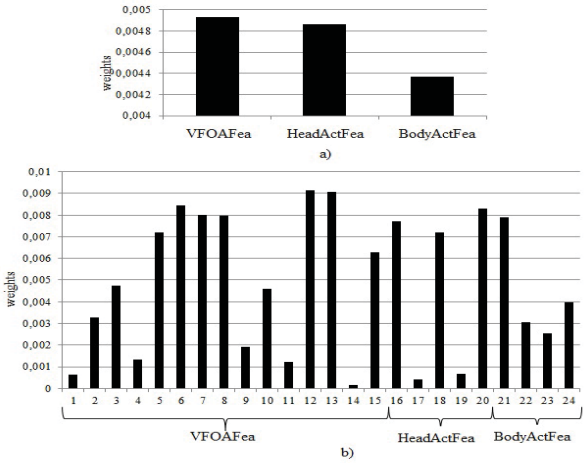

Figure 4: Average kernel weights for a) feature groups, b) each feature (ordered as in Table 4)

ActivityFea-SVM and BodyActivityFea-SVM. VFOAFea-RFLA performed much better than HeadActivityFea-RFLA and BodyActivityFea-RFLA while BodyActivityFea-RFLA performed better than HeadActivityFea-RFLA.

In general, the least EL detection performance of methods were worse than the most EL detection performance (Table 1). It is observed that EL prediction is much difficult task compared to the most EL prediction since the least EL and the rest class are overlapping. Similarly, the annotations of the least ELs were also more complicated compared to the most ELs as given in Section 3.2.

Similar conclusions were acquired when leave-one-meeting-segment-out was applied although the performance of methods were better.

The kernel weights obtained from LMKL can be used to extract the relative contributions of features when all features are concatenated [11, 12]. In this scope, important features have higher combination weights. The average kernel weights for each feature group and each feature are given in Figure 4 when 2 kernels were used. The weights show that different feature groups has similar importance while none of the feature is significantly more or less important than any other. In overall, totWatcher, minTwoWatcherNoME, THT and stdHx have smaller weights.

Table 4: Correlation Coefficient Values Between Questionnaires and Visual Nonverbal Features

Nonverbal 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
THL	0.22	0.05
THT	0.30	0.15
AHT	-0.23	-0.27
stdHx	-0.03	-0.01
stdHy	0.01	-0.04
TBL	0.48	0.37
TBT	0.53	0.41
ABT	0.12	0.14
stdB	0.40	0.39

6.2 Correlation Analysis

In Table 4, the correlation between variables derived from questionnaires and visual nonverbal features are given. Like in [6], the correlation analysis was performed when the meeting videos were evaluated as whole, rather than segmented.

As seen from Table 4, except totNoLook, lookSomeOne, stdInitiatorME, stdHx and stdHy, all other nonverbal features found correlated (nine of them had high correlation (values from 0.5 to 1.0 or -0.5 to -1.0), five of them had medium correlation (values from 0.3 to 0.5 or -0.3 to -0.5) and five of them had low correlation (values from 0.1 to 0.3 or -0.1 to -0.3)) with the results of SYMLOG-Observers. Similarly, except totNoLook, lookSomeOne, stdInitiatorME, totInterCurrME, THL, stdHx and stdHy, all other nonverbal features were correlated (seven of them had high correlation, six of them had medium correlation and four of them had low correlation) with the results of GLIS-Observer.

7. CONCLUSIONS

In this paper, different types of nonverbal features were combined to detect the ELs in a meeting environment. Different than the majority of the work in emergent leadership, only visual features were used. For the first time, gaze, head activity and body activity based features were combined which presented influential results showing that they are complementary for emergent leadership. The main novelty of this work is using LMKL which demonstrated improved results compared to the state of the art methods. Using LMKL, in average (multi-class and binary classifications together) 81% detection rate for the most EL and 73% detection rate for the least EL were obtained. It is also detected that the majority of the nonverbal features used were

highly correlated with the results of the social psychology questionnaires which test the leadership and the dominance.

When the meeting segments' labeling by human annotators and the results of SYMLOG-Observers and GLIS-Observers were compared using their highest/ lowest values for the most and the least EL inference, it has seen that there are a very high overlap. Such that, 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 were obtained.

As future work, audio nonverbal features will be extracted and combined with the existing visual nonverbal features, assuming that their combination might produce much better results. Additionally, the interactions between persons in a meeting scenario will be modeled as sequences of events using audio and video cues for identifying ELs and the co-occurrences of nonverbal features will be investigated.

8. 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] O. Aran and D. Gatica-Perez. One of a kind: inferring personality impressions in meetings. In *ICMI*, pages 9–13, 2013.

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

[5] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt. Performance of optical flow techniques. In *IEEE CVPR*, pages 236–242, 1992.

[6] 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 *ICMI*, 2016.

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

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

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

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

[11] M. Gonen and E. Alpaydin. Localized multiple kernel learning. In *ICML*, pages 352–359, 2008.

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

[13] C. Hsu and C. Lin. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Networks*, 13:415–425, 2002.

[14] 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.

[15] 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.

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

[17] A. Kindiroglu, L. Akarun, and O. Aran. Vision based personality analysis using transfer learning methods. In *IEEE SIU*, pages 2058–2061, April 2014.

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

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

[20] 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.

[21] L. McDowell, K. Gupta, and D. Aha. Cautious collective classification. *Journal of Machine Learning Research*, 10:2777–2836, 2009.

[22] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Trans. on Multimedia*, 16(4):1018–1031, June 2014.

[23] V. Paul and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, pages 511–518, 2001.

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

[25] 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.

[26] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez. Identifying emergent leadership in small groups using nonverbal communicative cues. pages 8–10. ICMI-MLMI, 2010.

[27] 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.

[28] 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.

[29] M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *ICML,*, 2009.

[30] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

[31] 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.