Dominance Detection in Face-to-face Conversations

[1,3]Sergio Escalera, [2]Rosa M. Martínez, [1,3]Jordi Vitrià, [1,3]Petia Radeva, and [2]M. Teresa Anguera
[1]Dept. Matemàtiques, Universitat de Barcelona, Gran Via de les Corts, 585, 08007, Barcelona, Spain.
[2]Dept. Metodologia de les Ciències del Comportament, Universitat de Barcelona,
Gran Via de les Corts, 585, 08007, Barcelona, Spain.
[3]Computer Vision Center, Campus UAB, edifici O, 08193, Bellaterra, Barcelona, Spain.
sergio@maia.ub.es, rosamaria_mf@hotmail.com, {jordi,petia}@maia.ub.es, mtanguera@gmail.com

Abstract

Dominance is referred to the level of influence a person has in a conversation. Dominance is an important research area in social psychology, but the problem of its automatic estimation is a very recent topic in the contexts of social and wearable computing. In this paper, we focus on dominance detection from visual cues. We estimate the correlation among observers by categorizing the dominant people in a set of face-to-face conversations. Different dominance indicators from gestural communication are defined, manually annotated, and compared to the observers opinion. Moreover, the considered indicators are automatically extracted from video sequences and learnt by using binary classifiers. Results from the three analysis shows a high correlation and allows the categorization of dominant people in public discussion video sequences.

1. Introduction

Four of the most well-known activities studied in group conversations are: addressing, turn-taking, interest, and dominance or influence [1]. Addressing refers to whom the speech is directed. Turn-taking patterns in group meetings can be potentially used to distinguish several situations, such as monologues, discussions, presentations, note-taking, etc [2]. The group interest can be defined as the degree of engagement that the members of a group collectively display during their interaction. Finally, dominance is concerned to the capability of a speaker to drive the conversation and to have large influence on the meeting. Although dominance is an important research area in social psychology [3], the problem of its automatic estimation is a very recent topic in the context of social and wearable computing [4, 5, 6, 7]. Dominance is often seen in two ways, both "as a personality characteristic" (a trait) and "to indicate a persons hierarchical position within a group (a state). Although dominance and related terms like power have multiple definitions and are often used as equivalent, a distinguishing approach defines power as "the capacity to produce intended effects, and in particular, the ability to influence the behavior of another person" [8].

In this paper, we focus on the recognition of dominant people as a state in face-to-face conversations. State-of-the-art studies for dominance detection generally work with visual and audio cues in group meetings. For example, Rienks et al. [4] proposed a supervised learning approach to detect dominance in meetings based on the formulation of a manually-annotated three-class problem, consisting of high, normal, and low dominance classes. Related works [6, 7] use features related to speaker-turns, speech transcriptions, or addressing labels. Also, people status and look have shown to be dominance indicators [9]. Most of these works define a conversational environment with several participants, and dominance and other indicators are quantified using pair-wise measurements and rating the final estimations. However, the automatic estimation of dominance and the relevant cues for its computation remain as an open research problem.

In this paper, we focus on gestural communication in face-to-face interactions. We selected a set of dyadic discussions from a public video dataset depicting face to face interactions in the New York Times web site [10]. The conversations were shown to several observers that labeled the dominance based on their personal opinion. Speaking time, stress, visual focus, and successful interruptions were defined, manually annotated, and automatically extracted. We omitted the audio cues in order to determine the influence of visual cues in the dominance detection problem. The three analysis: observers opinion, manually annotated indicators, and automatic feature extraction and classification, shown statistically significant correlation discriminating among dominant and dominated people.

The paper is organized as follows: Section 2 presents the definition and computation of the visual cues for dominance

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

detection. Section 3 describes the experimental validation by means of observers labeling, indicator manual annotation, and automatic feature extraction and dominance classification. Finally, Section 4 concludes the paper.

2. Dominance indicators

In order to detect the *dominant* person in a face-to-face interaction video sequence, we must first define a set of basic visual features. These features are based on the movement of the individual subjects. Then, a post-processing is applied in order to regularize the motion feature vectors. Finally, the motion feature vectors serve as bases to build the higher level dominance indicators.

2.1. Motion-based basic features

Given a video sequence $S = \{s_1, .., s_e\}$, where s_i is the ith frame in a sequence of e frames with a resolution of $h \times w$ pixels, we define three individual signal features: global motion, face motion, and mouth motion. Given two frames s_i and s_j, the corresponding global motion GM_{ij} is estimated as the accumulated sum of the absolute value of the subtraction between two frames, s_i and s_j.

Since the faces that appear in our dialog sequences are almost all of them in frontal view, we can make use of the state-of-the-art face detectors to compute the face movement. In particular we use the Viola & Jones face detector, and compute the face motion feature FM_{ij} at ith frame as $FM_{ij} = \frac{1}{n \cdot m} \sum_k |F_{j,k} - F_{i,k}|$, where $F_{i,k}$ is the kth pixel in face region F_i, $k \in \{1, .., n \cdot m\}$, and term $n \cdot m$ normalizes the face motion feature. Since some faces can be non-detected because of fast head movements, the face detected at the previous frame is also considered as the new one. On the other hand, in case of having false positives (that is, more than one detected face by speaker), the one with the minimum distance in respect to the face detected at the previous frame is the selected one.

Finally, we compute the mouth motion MM_{il} at frame i. For this task, we estimate an accumulated substraction of l mouth regions previous to the mouth at frame i. From the face region $F_i \in \{0, .., 255\}^{n \times m}$ detected at frame i, the mouth region is defined as $M_i \in \{0, .., 255\}^{n/2 \times m/2}$, which corresponds to the center bottom half region of F_i. Then, given the parameter l, the mouth motion feature MM_{il} is computed as $MM_{il} = \frac{1}{n \cdot m/4} \sum_{j=i-l}^{i-1} \sum_k |M_{i,k} - M_{j,k}|$, where $M_{i,k}$ is the kth pixel in a mouth region M_i, $k \in \{1, .., n \cdot m/4\}$, and $n \cdot m/4$ is a normalizing factor.

2.2. Post-processing

After computing the values of GM_{ij}, FM_{ij}, and MM_{il} for a sequence of e frames ($i, j \in [1, .., e]$), we obtain their corresponding motion-based vectors. At the post-processing step, first, we filter the vectors in order to obtain a 3-value quantification. For this task, all vectors from all speakers for each movement feature are considered together to compute the corresponding feature histogram (i.e. histogram of global motion h_{GM}), which is normalized to unit in order to estimate the probability density function (i.e. pdf of global motion P_{GM}). Then, two thresholds are computed in order to define the three values of motion, corresponding to low, medium, and high motion quantifications:

$$t_1 : \int_0^{t_1} P_{GM} = \frac{1}{3}, \quad t_2 : \int_0^{t_2} P_{GM} = \frac{2}{3} \qquad (1)$$

Finally, in order to avoid abrupt changes in short sequences of frames, we apply a sliding window filtering of size q using a majority voting rule. The result of this step is a smoother vector V (i.e. vector of global motion V_{GM}).

2.3. Dominance-based features

Most of the state-of-the-art works related to dominance detection are focused on verbal cues in group meetings. In this work we focus on non-verbal cues in face-to-face interactions. In this sense, we defined the following set of visual dominance features:

• **Speaking Time - ST**: We consider the time a participant is speaking in the meeting as an indicator of dominance.

• **The number of successful interruptions - NSI**: The number of times a participant interrupts to another participant making him stop speaking is an indicator of dominance.

• **The number of times the floor is grabbed by a participant - NOF**: When a participant grabs the floor is an indicator of being dominated.

• **The speaker gesticulation degree - SGD**: Some studies suggest that high degree of gesticulation of a participant when speaking makes the rest of participants to focus on him, being a possible indicator of dominance (also known as stress [11]).

There are several other indicators of dominance, such as the influence diffusion, addressing, turn-taking, number of questions, etc. However, most of them require audio features, or several participants and ranking features. In this work, we want to analyze if the previous simple non-verbal cues have enough discriminability power to generalize the dominance in the face-to-face conversational data analyzed in this paper.

Next, we describe how we compute these dominance features using the simple motion-based non-verbal cues presented in the previous section.

We can compute the speaking time ST based on the degree of participant mouth movement during the meeting as

follows:

$$ST^1 = \frac{\sum_{i=1}^{k} V^1_{MM_i}}{\max(\sum_{i=1}^{k} V^1_{MM_i} + \sum_{i=1}^{k} V^2_{MM_i}, 1)}, \quad ST^2 = 1 - ST^1 \quad (2)$$

where ST^1 and ST^2 stand for the percentage of speaking time $\in [0, .., 1]$ during conversation of participants 1 and 2, respectively.

Given the 3-value mouth motion vectors V^1_{MM} and V^2_{MM} for both participants, we define a successful interruption I^2 of the second participant if the following constraint is satisfied:

$$V^{1,2}_{MM_{i-1}} = 0, \quad V^{1,2}_{MM_i} = 1, \sum_{j=1-z}^{i} V^2_{MM_j} < \frac{z}{2}, \quad (3)$$

$$\sum_{j=i}^{i+z} V^2_{MM_j} > \frac{z}{2}, \quad \sum_{j=1-z}^{i} V^1_{MM_j} > \frac{z}{2}, \quad \sum_{j=i}^{i+z} V^1_{MM_j} < \frac{z}{2} \quad (4)$$

where we consider a width of z frames to analyze the interruption and $V^{1,2}_{MM_i}$ is computed as $V^{1,2}_{MM_i}$=$V^1_{MM_i} \cdot V^2_{MM_i}$. An example of a successful interruption I^2 of the second speaker is shown in Figure 1.

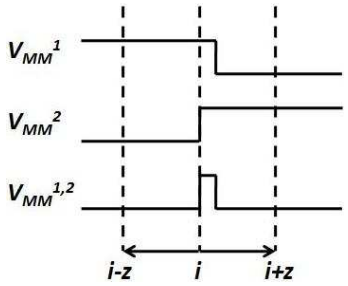

Figure 1. Interruption measurement.

Then, the percentage of successful interruption by a participant is defined as follows:

$$NSI^1 = \frac{|I^1|}{\max(|I^1| + |I^2|, 1)}, \quad NSI^2 = 1 - NSI^1 \quad (5)$$

where $|I^i|$ stands for the number of successful interruptions of the ith participant.

We approximate the number of times the floor is grabbed by a participant (NOF) as the amount downward motion executed by that particiapant. This feature can be approximated by the magnitude of the derivative of the sequence of frames respect to the time $|\frac{\partial S}{\partial t}|$, which codifies the motion produced between consecutive frames. In order to obtain the vertical movement orientation to approximate the NOF feature, we compute the derivative in time of the previous measurement as $\frac{\partial |\frac{\partial S}{\partial t}|}{\partial t}$. Figure 2 shows the two derivatives for an input sequence. The blue regions marked in the last image correspond to the highest changes in orientation. In order to compute the derivative orientation, we estimate the number of changes from positive to negative and negative to positive in the vertical direction from up to down in the image. Then, the magnitude of the derivative $\sum \frac{\partial |\frac{\partial S}{\partial t}|}{\partial t}$ is used in positive for down orientations or negative for up orientations. This feature vector VM^i codifies the i-user face movement in the vertical axis.

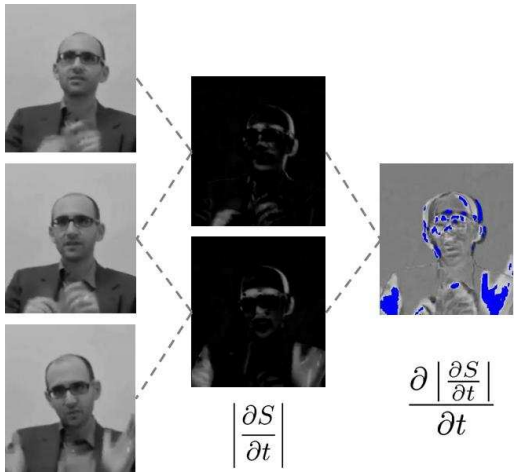

Figure 2. Vertical movement approximation.

Finally, the NOF feature is computed as follows:

$$NOF^1 = \frac{\sum_i VM^1_i}{\max(\sum_i VM^1_i + \sum_i VM^2_i, 1)}, \quad NOF^2 = 1 - NOF^1 \quad (6)$$

The speaker gesticulation degree SGD refers to the variation in emphasis. We compute this feature as follows:

$$\forall k \in \{1, .., e\}, V^i_{MM_k} := \min(1, V^i_{MM_k}) \quad (7)$$

$$G = \left(V^i_{MM} \cdot V^i_{GM}\right) / \sum_k V^i_{MM_k} \quad (8)$$

where $i \in \{1, 2\}$ is the speaker and $k \in \{1, .., e\}$. This measure corresponds to the global motion of each person, only taking into account the time when he is speaking, and normalizing this value by the speaking time. This feature is computed for each speaker separately (G_1 and G_2). Finally, the SGD feature is defined as follows:

$$SGD^1 = \frac{\sum_i G^1_i}{\max(\sum_i G^1_i + \sum_i G^2_i, 1)}, \quad SGD^2 = 1 - SGD^1 \quad (9)$$

3. Experiments and Results

In order to present the results, first we discuss the data, methods, and experiments.

• *Data*: The data used for the experiments consists of dyadic video sequences from the public New York Times web site video library [10]. In each conversation, two

speakers with different points of view discuss about a direct question (i.e. "In the fight against terrorism, is an American victory in sight?"). From this data set, seven videos have been selected. These videos are shown in Figure 3. To compare videos at similar conditions, all speakers are midage men. Each video has a frame rate of 12 FPS and a duration of four minutes, which correspond to 2880 frames video sequences.

Figure 3. Blogging heads face-to-face conversations.

• $Methods$: In order to train a binary classifier to learn the dominance features we have used different classifiers: Discrete Adaboost with decision stumps [12], Linear Support Vector Machines with the regularization parameter $C = 1$ [13], Support Vector Machines with Radial Basis Function kernel with $C = 1$ and $\sigma = 0.5$ [13], Fisher Linear Discriminant Analysis using 99% of the principal components [14], and Nearest Mean Classifier.

• $Experiments$: First, we asked 40 independent observers to put a label on each of the videos. Observers were not aware of the objective of the experiment. After looking for the correlation of dominance labels among observers answers, we manually and automatically annotated and computed the ST, NSI, NOF, and SGD dominance indicators, and analyzed them to look for their relation to the observers opinion. Finally, we performed the same procedure using the automatic feature extraction methodology.

3.1. Observers inquiry

We performed an experiment with 40 people from 13 different nationalities asking for their opinion regarding the most dominant people in the seven New York Times dyadic conversations. The observers labelled each dominant people for each conversation, only taking into account the visual information (omitting audio), based on their personal notion of dominance. Since each video is composed of a left and a right speaker, we labeled the left dominance opinions as one and the right dominant decisions as two. In order to compare the correlation among observers, the mean value for each video using the 40 opinions is computed. Thus, values near one or values near two correspond to high correlation among observers opinion deciding the most dominant participant as the left or the right participant, respectively. In this way, we computed the confidence P of the correlation C of each video as follows:

$$P = 1 - \min(2 - C, C - 1) \tag{10}$$

The results of P based on the observers opinion are numerically shown for each video in Figure 4. Note that all results are in the range [65,..,95] of confidence, which corresponds to high correlation among observers opinion.

3.2. Labeled data

In order to analyze if the dominance indicators defined at the previous sections have discriminative power to obtain similar results to those reported by observers, we manually annotated the indicators over the dyadic video sequences. For each four minutes video sequence, intervals of ten seconds are defined for each participant. This corresponds to 24 intervals for four indicators and two participants, with a total of 192 manually annotated values per video sequence (1344 manual values considering the set of seven videos). The indicators correspond to speaking, successful interruption, grab the floor, and gesticulate while speaking, respectively. If an indicator appears within an interval of ten seconds, the indicator is activated for that participant and that interval independently of the time the indicator appears.

In order to manually fill the indicators, three different people annotated the video sequences, and the value of each indicator position is set to one if the majority from the three labelers activate the indicator or zero otherwise. After the manual labeling, for each dyadic conversation, the ST, NSI, NOF, and SGD dominance features are computed by summing the values of the indicators and computing its percentage as defined in equations (2), (5), (6), and (9), respectively. The numerical results are graphically shown in the blue bars of graphics in Figures 5(a)-(g) for the seven dyadic sequences, respectively. Using the observers criterion, the indicators values of the dominant speakers are shown in the left of the graphics, and the dominated participants in the right part of the graphics, respectively.

In order to determine if the computed values for the indicators generalize the observers opinion, we performed a binary classification experiment. We used Adaboost in a set of leave-one-out experiments. Each experiment uses one iteration of decision stumps over a different dominance indicator. Classification results are shown in Table 1. Note that all indicators attain classification accuracy upon 70% based on the groups of classes defined by the observers. Moreover, the ST indicator is able to classify all the videos as expected by the observers.

Indicator	Accuracy
Manual ST	100 %
Manual NSI	86 %
Manual NOF	71 %
Manual SGD	71 %

Table 1. Dominance classification results using independent manually-labeled indicators.

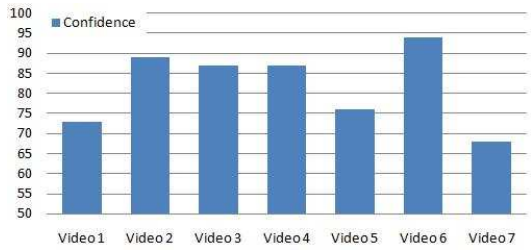

Figure 4. Observers correlation values.

3.3. Automatic dominance features

For this experiment, we automatically computed the ST, NSI, NOF, and SGD dominance indicators as explained in the previous section. The videos are in 12FPS, and four minutes per video defines independent sequences of 2880 frames, representing a total of 20160 analyzed frames. The mouth history in frames and the windows size for the successful interruption computation are set to ten. The numerical values obtained are shown in the red bars of Figures 3(b)-(h) next to the manual results of the previous experiment. Note that the obtained results are very similar to the percentages obtained by the manual labeling. Next, we perform a binary classification experiment to analyze if the new classification results are also maintained respect to the previous manual labeling. The performance results applying a leave-on-out experiment over each feature using one decision stump of Adaboost are shown in Table 2. Note that except in the case of the NSI indicator, which slightly reduces the performance in the case of the automatic features, the rest of performance results are maintained for the remaining indicators.

Finally, in order to analyze the whole set of dominance indicators together to solve the dominant detection problem, we used a set of classifiers, performing two experiments. The first experiment corresponds to a leave-one-out

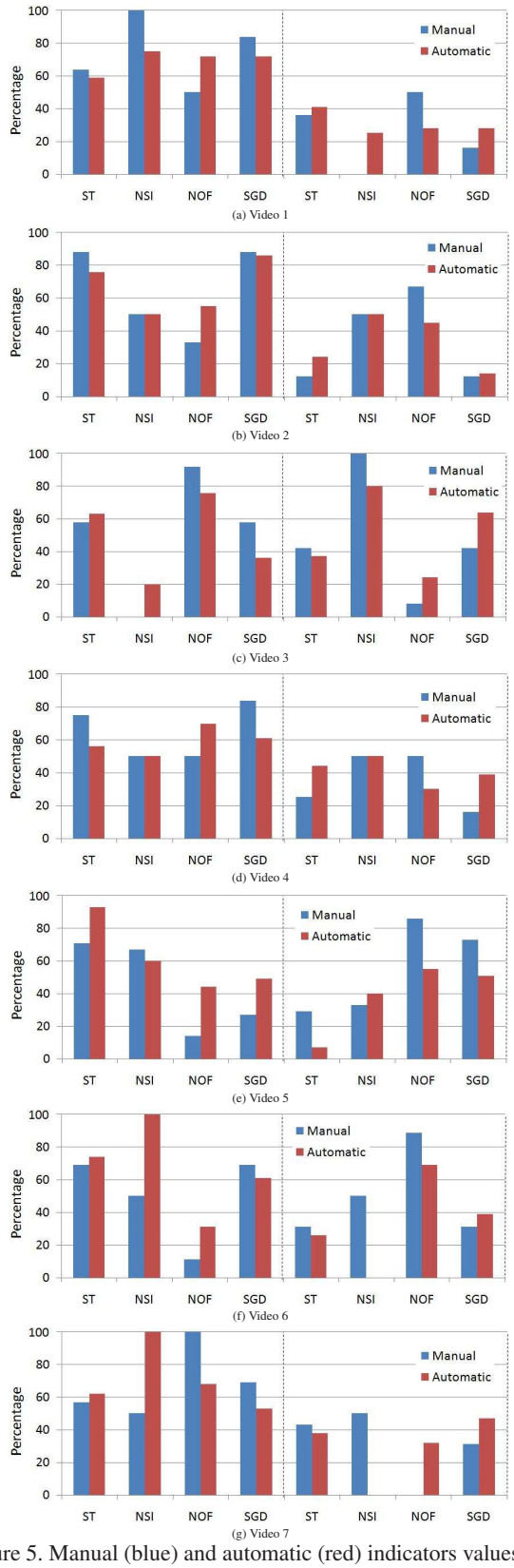

Figure 5. Manual (blue) and automatic (red) indicators values.

Indicator	Accuracy
Automatic ST	100 %
Automatic NSI	79 %
Automatic NOF	71 %
Automatic SGD	71 %

Table 2. Dominance classification results using independent automatic-extracted dominance indicators.

evaluation, and the second one to a bootstrap [15] evaluation. To perform a bootstrap evaluation, 200 random sequences of videos were defined, where each sequence has seven possible values, each one corresponding to the label of a possible video randomly selected. Then, to evaluate the performance over each video, all sequences which do not consider the video are selected, and using the indicated videos in the sequence a binary classifier splitting dominant and dominated participant classes is learnt and tested over the omitted video. After computing the seven performances for the seven videos, the mean accuracy corresponds to the global performance. Note that this evaluation strategy is more pessimistic since based on the random sequences different number of videos are used to learn the classifier, and thus, generalization becomes more difficult to achieve by the classifier. The classification results in the case of the leave-one-out and bootstrap evaluations are shown in Table 3. The results in the case of the leave-one-out evaluation show high accuracy predicting the dominance criterion of observers for all types of classifiers, slightly reducing the performance in the case of Linear SVM and NMC. The results for the bootstrap evaluation are in general lower than at the leave-one-out experiment. However, except in the case of the NMC, all classifiers obtain results around 90% of accuracy.

Learning strategy	Accuracy	Accuracy
Discrete Adaboost	100 %	93.62 %
Linear SVM	85.71 %	88.82 %
RBF SVM	100 %	86.83 %
FLDA	100 %	91.28 %
NMC	85.71 %	76.90 %

Table 3. Dominance classification results using dominance indicators and leave-one-out evaluation (first column) and bootstrap evaluation (second column).

4. Conclusions

We analyzed a set of non-verbal cues to detect the dominant people in face-to-face video sequences from the New York Times web site. We performed an experiment with 40 observers asking for their opinion regarding the most influent participant in a set of dyadic sequences. Results shown high correlation among observers opinion. We also defined a set of gestural communication indicators and manually annotated the videos. Comparing to the observers opinion, the indicators shown high discriminative power. Moreover, an automatic approximation to the dominant features based on low-level movement-based features was presented. Results shown high correlation among dominance prediction for three: observers, manually annotated, and automatic approach.

5. Acknowledgements

This work has been supported in part by projects TIN2006-15308-C02, FIS PI061290, and CONSOLIDER-INGENIO CSD 2007-00018.

References

[1] D. Gatica-Perez, Analyzing group interactions in conversations: a review, in: Multisensor Fusion and Integration for Intelligent Systems, 2006, pp. 41–46. 1

[2] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, D. Zhang, Automatic analysis of multimodal group actions in meetings, in: PAMI, Vol. 27, 2005, pp. 305–317. 1

[3] S. L. Ellyson, J. F. Dovidio, Power, dominance, and nonverbal behavior, in: Springer-Verlag, 1985. 1

[4] R. Rienks, D. Heylen, Automatic dominance detection in meetings using svm, in: MLMI, 2005. 1

[5] D. Babu, H. Hung, C. Yeo, D. Gatica-Perez, Modeling dominance in group conversations using nonverbal activity cues, in: Audio, Speech, and Language Processing, IEEE Transactions on, Vol. 17, 2009, pp. 501–513. 1

[6] H. Hung, D. Babu, S. Ba, J. Odobez, D. Gatica-Perez, Investigating automatic dominance estimation in groups from visual attention and speaking activity, in: International Conference on Multimodal Interfaces, 2008, pp. 233–236. 1

[7] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J. Odobez, K. Ramchandran, N. Mirghafori, D. Gatica-Perez, Using audio and video features to classify the most dominant person in a group meeting, in: International Multimedia Conference, 2007, pp. 835–838. 1

[8] D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: a review, in: Image and Vision Computing, 2009. 1

[9] J. S. Efran, Looking for approval: effects of visual behavior of approbation from persons differing in importance, in: Journal of Personality and Social Psychology, Vol. 10, 1968, pp. 21–25. 1

[10] http://video.nytimes.com/. 1, 3

[11] A. Pentland, Socially aware computation and communication, in: Computer, Vol. 38, 2005, pp. 33–40. 2

[12] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, in: The annals of statistics, Vol. 38, 1998, pp. 337–374. 4

[13] Osu-svm-toolbox.
URL http://svm.sourceforge.net 4

[14] P. Tools, http://prtools.org/. 4

[15] B. Efron, R. Tibshirani, An introduction to the bootstrap, in: Monographs on Statistics and Applied Probability, 1993. 6