Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 491819, 12 pages
doi:10.1155/2010/491819

Research Article

Automatic Detection of Dominance and Expected Interest

Sergio Escalera,[1,2] Oriol Pujol,[1,2] Petia Radeva,[1,2] Jordi Vitrià,[1,2] and M. Teresa Anguera[3]

[1] *Computer Vision Center, Campus UAB, Edifici O, 08193 Bellaterra, Spain*
[2] *Departament de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain*
[3] *Departament de Metodologia de les Ciències del Comportament, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain*

Correspondence should be addressed to Sergio Escalera, sescalera@cvc.uab.es

Received 3 August 2009; Revised 24 December 2009; Accepted 17 March 2010

Academic Editor: Satya Dharanipragada

Social Signal Processing is an emergent area of research that focuses on the analysis of social constructs. Dominance and interest are two of these social constructs. Dominance refers to the level of influence a person has in a conversation. Interest, when referred in terms of group interactions, can be defined as the degree of engagement that the members of a group collectively display during their interaction. In this paper, we argue that only using behavioral motion information, we are able to predict the interest of observers when looking at face-to-face interactions as well as the dominant people. First, we propose a simple set of movement-based features from body, face, and mouth activity in order to define a higher set of interaction indicators. The considered indicators are manually annotated by observers. Based on the opinions obtained, we define an automatic binary dominance detection problem and a multiclass interest quantification problem. Error-Correcting Output Codes framework is used to learn to rank the perceived observer's interest in face-to-face interactions meanwhile Adaboost is used to solve the dominant detection problem. The automatic system shows good correlation between the automatic categorization results and the manual ranking made by the observers in both dominance and interest detection problems.

1. Introduction

For most of us, social perception is used unconsciously for some of the most important actions we take in our life: negotiating economic and affective resources, making new friends, and establishing credibility, or leadership. Social Signal Processing [1] and Affective Computing [2–4] are emergent areas of research that focus on the analysis of *social cues* and personal traits [5–7]. The basic signals come from different sources and include gestures, such as scratching, head nods, *huh* utterances, or facial expressions. As such, automatic systems in this line of work benefit of technologies such as face detection and localization, head and face tracking, facial expression analysis, body detection and tracking, visual analysis of body gestures, posture recognition, activity recognition, estimation of audio features such as pitch, intensity, and speech rate, and the recognition of nonlinguistic vocalizations like laughs, cries, sighs, and coughs [8]. However, humans group these basic signals to form social messages (i.e., dominance, trustworthiness, friendliness, etc.), which take place in group interactions.

Four of the most well-known studied group activities in conversations are: addressing, turn-taking, interest, and dominance or influence [9]. 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, and note-taking [10]. 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 [11], the problem of its automatic estimation is a very recent topic in the context of social and wearable computing [12–15]. Dominance is often seen in two ways, both "as a personality characteristic" (a trait) and to indicate a person's 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" [16].

Concerting the term interest, it is often used to designate people's internal states related to the degree of engagement that individuals display, consciously or not, during their interaction. Such displayed engagement can be the result of many factors, ranging from interest in a conversation, attraction to the interlocutor(s), and social rapport [17]. In the specific context of group interaction, the degree of interest that the members of a group collectively display during their interaction is an important state to extract from formal meetings and other conversational settings. Segments of conversations where participants are highly engaged (e.g., in a discussion) are likely to be of interest to other observers too [17].

Most of the studies in dominance and interest detection generally work with visual and audio cues in group meetings. For example,Rienks and Heylen [12] 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 [14, 15] use features related to speaker-turns, speech transcriptions, or addressing labels. Also, people status and look have shown to be dominance indicators [18]. 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 the case of interest, the authors of [19, 20] proposed a small set of social signals, such as activity level, stress, speaking engagement, and corporal engagement for analyzing nonverbal *speech* patterns during dyadic interactions.

In this article, we give an approximation to the quantification of dominance and perceived interest from the point of view of an external observer exclusively analyzing visual cues. Note that, contrary to many studies that pursue the assessment of participants' interest and use them as a surrogate feature to assess observer's own interest [21], this article directly addresses perceived observer's interest in face-to-face interactions.[1] In particular, our approach focuses 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 [22]. The conversations were shown to several observers that labeled the dominance and interest based on their personal opinion, defining the groundtruth data. We argue that only using behavioural motion information, we are able to predict the perceived dominance and interest by observers. From the computation of a set of simple motion-based features, we defined a higher set of interaction features: speaking time, stress, visual focus, and successful interruptions for dominance detection, and stress, activity, speaking engagement, and corporal engagement for perceived interest quantification. These features are learnt with Adaboost and the Error-Correcting Output Codes framework to obtain a dominance detection and interest quantification methodologies. Three analyses: observers opinion, manually annotated indicators, and automatic feature extraction and classification show statistically significant correlation discriminating among dominant-dominated people and ranking the observer's level of interest.

The layout of the article is as follows: Section 2 presents the motion-based features and the design of the dominance and interest indicators. Section 3 reviews the machine learning framework used in the paper. Section 4 describes the experimental validation by means of observers labeling, indicator manual annotation, and automatic feature extraction and classification. Finally, Section 5 concludes the paper.

2. Dominance and Interest Indicators

In order to predict dominant people and the level of interest perceived by observers when looking at face-to-face interactions, first, we define a set of basic visual features. These features are based on the movement of the individual subjects. Then, a postprocessing is applied in order to regularize the movement features. These features will serve as bases to build higher level interaction features, commonly named as indicators in psychology, for describing the dominance and interest constructs.

2.1. Movement-Based Basic Features. Given a video sequence $S = \{s_1, \ldots, s_e\}$, where s_i is the ith frame in a sequence of e frames with a resolution of $h \times w$ pixels, we define four individual signal features: global movement, face movement, body movement, and mouth movement.

(i) Global Movement. Given two frames s_i and s_j, the global movement GM_{ij} is estimated as the accumulated sum of the absolute value of the subtraction between two frames s_i and s_j:

$$\mathrm{GM}_{ij} = \sum_k \left| s_{j,k} - s_{i,k} \right|, \tag{1}$$

where $s_{i,k}$ is the kth pixel in frame s_i, $k \in \{1, \ldots, h \cdot w\}$. Figure 1(a) shows a frame from a dialog, and Figure 1(b) its corresponding GM_{ij} image, where i and j are consecutive frames in a 12 FPS video sequence.

(ii) Face Movement. 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. In particular, the face detector of Viola and Jones [23] is one of the most widely applied detectors due to its fast computation and high detection accuracy, at the same time that it preserves a low false alarm rate. We use the face detector trained using a Gentle version of Adaboost with decision stumps [23]. The Haar-like features and the rotated ones have been used to define the feature space [23]. Figure 1(c) shows an example

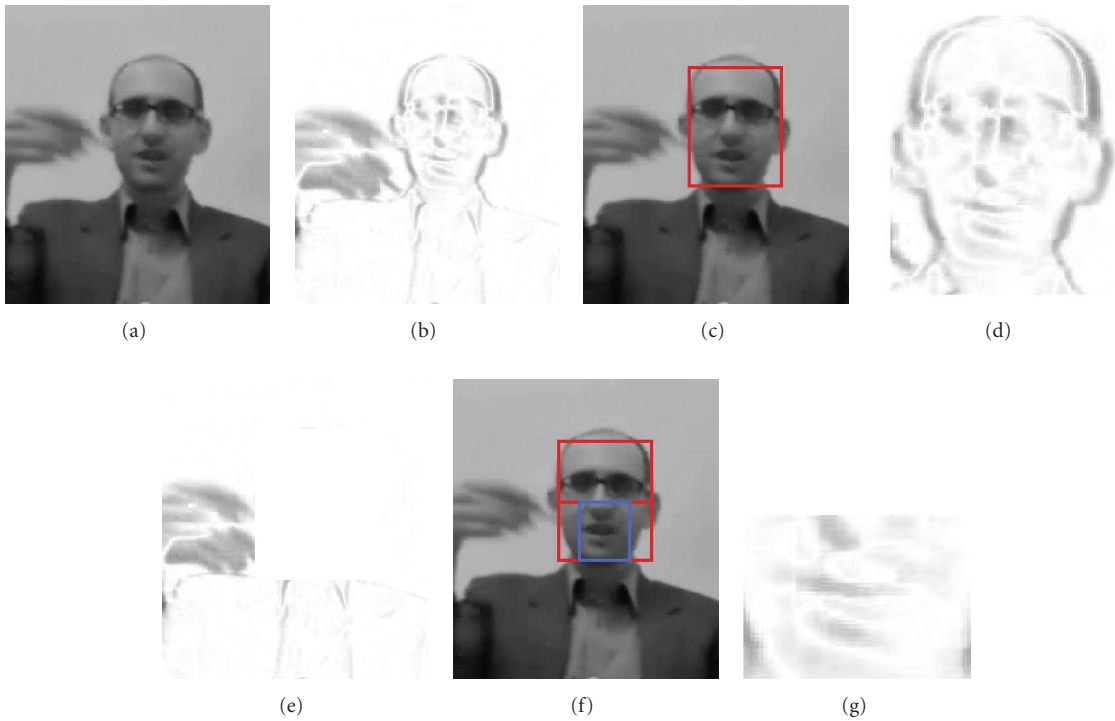

Figure 1: (a) ith frame from dialog, (b) Global movement GM_{ij}, (c) Detected face F_i, (d) Face movement FM_{ij}, (e) Body movement BM_{ij}, (f) Mouth detection M_i, and (g) Mouth movement MM_{ij}.

of a detected face of size $n \times m$, in the ith frame of a sequence, denoted by $F_i \in \{0,\ldots,255\}^{n\times m}$. Then, the face movement feature FM_{ij} at ith frame is defined as follows:

$$\mathrm{FM}_{ij} = \frac{1}{n \cdot m} \sum_{k} \left| F_{j,k} - F_{i,k} \right|, \tag{2}$$

where $F_{i,k}$ is the kth pixel in face region F_i, $k \in \{1,\ldots,n \cdot m\}$, and the term $n \cdot m$ normalizes the face movement feature. An example of faces substraction $|F_j - F_i|$ is shown in Figure 1(d).

(iii) Body Movement. We define the body movement BM as follows:

$$\mathrm{BM}_{ij} = \sum_{k} \left| s_{i,k} - s_{j,k} \right| - \sum_{f_k \in F^{ij}} f_k. \tag{3}$$

In this case, the pixels f_k corresponding to the bounding box F^{ij} which contains both faces F_i and F_j are removed from the set of pixels that defines the global movement image of frame i. An example of a body image substraction is shown in Figure 1(e).

(iv) Mouth Movement. In order to avoid the bias that can appear due to the translation of mouth detection between consecutive frames, for computing the mouth movement MM_{iL} at frame i, we estimate an accumulated substraction of L mouth regions previous to the mouth at frame i. From the face region $F_i \in \{0,\ldots,255\}^{n\times m}$ detected at frame i, the mouth region is defined as $M_i \in \{0,\ldots,255\}^{n/2\times m/2}$, which corresponds to the center bottom half region of F_i. Then, given the parameter L, the mouth movement feature MM_{iL} is computed as follows:

$$\mathrm{MM}_{iL} = \frac{1}{n \cdot m/4} \sum_{j=i-L}^{i-1} \sum_{k} \left| M_{i,k} - M_{j,k} \right|, \tag{4}$$

where $M_{i,k}$ is the kth pixel in a mouth region M_i, $k \in \{1,\ldots,n \cdot m/4\}$, and $n \cdot m/4$ is a normalizing factor. The accumulated subtraction avoids false positive mouth activity detection due to noisy data and translation artifacts of the mouth region. An example of a detected mouth F_i is shown in Figure 1(f), and its corresponding accumulated substraction for $L = 3$ is shown in Figure 1(g).

2.2. Post-Processing. After computing the values of GM_{ij}, FM_{ij}, BM_{ij}, and MM_{iL} for a sequence of e frames ($i, j \in [1,\ldots,e]$), we filter the responses. Figures 2(c) and 2(d) correspond to the global movement features GM_{ij} in a sequence of 5000 frames at 12 FPS for the speakers of Figures 2(a) and 2(b), respectively. At the post-processing step, first, we filter the features in order to obtain a 3-value quantification. For this task, all feature values from all speakers for each movement feature are considered together to compute the corresponding feature histogram (i.e., histogram of global movement h_{GM}), which is normalized to estimate the probability density function (i.e., pdf of global movement P_{GM}). Then, two thresholds are computed in

order to define the three values of movement, corresponding to low, medium, and high movement quantifications:

$$t_1 : \int_0^{t_1} P_{\mathrm{GM}} = \frac{1}{3}, \qquad t_2 : \int_0^{t_2} P_{\mathrm{GM}} = \frac{2}{3}. \tag{5}$$

The result of this step is shown in Figures 2(e) and 2(f), respectively.

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 smooth result of this step is denoted by V (Figures 2(g) and 2(h), resp.).

2.3. Dominance-Based Indicators. 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 nonverbal cues in face-to-face interactions. In this sense, we defined the following set of visual dominance features.

(i) Speaking Time or Activity—ST. We consider the time a participant is speaking in the meeting as an indicator of dominance.

(ii) 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.

(iii) The Number of Times the Floor Is Grabbed by a Participant—NOF. When a participant grabs the floor is an indicator of being dominated.

(iv) 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 [19]).

There are several other indicators of dominance, such as the influence diffusion, addressing, turn-taking, and number of questions. 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:

$$\mathrm{ST}^1 = \frac{\sum_{i=1}^{k} V^1_{\mathrm{MM}_i}}{\max\left(\sum_{i=1}^{k} V^1_{\mathrm{MM}_i} + \sum_{i=1}^{k} V^2_{\mathrm{MM}_i}, 1\right)}, \quad \mathrm{ST}^2 = 1 - \mathrm{ST}^1, \tag{6}$$

where ST^1 and ST^2 stand for the percentage of speaking time $\in [0, \ldots, 1]$ 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:

$$V^{1,2}_{\mathrm{MM}_{i-1}} = 0, \quad V^{1,2}_{\mathrm{MM}_i} = 1, \quad \sum_{j=1-z}^{i} V^2_{\mathrm{MM}_j} < \frac{z}{2},$$
$$\sum_{j=i}^{i+z} V^2_{\mathrm{MM}_j} > \frac{z}{2}, \quad \sum_{j=1-z}^{i} V^1_{\mathrm{MM}_j} > \frac{z}{2}, \quad \sum_{j=i}^{i+z} V^1_{\mathrm{MM}_j} < \frac{z}{2}, \tag{7}$$

where we consider a width of z frames to analyze the interruption and $V^{1,2}_{\mathrm{MM}_i}$ is computed as $V^{1,2}_{\mathrm{MM}_i} = V^1_{\mathrm{MM}_i} \cdot V^2_{\mathrm{MM}_i}$. An example of a successful interruption I^2 of the second speaker is shown in Figure 3.

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

$$\mathrm{NSI}^1 = \frac{|I^1|}{\max(|I^1| + |I^2|, 1)}, \qquad \mathrm{NSI}^2 = 1 - \mathrm{NSI}^1, \tag{8}$$

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

In the case of the number of times the floor is grabbed by a participant (NOF), we can approximate this feature looking for downward movements of the participants. If the participant is detected in frontal view and then a downward movement occurs, it is straightforward to conclude that the participant is looking at the floor. In this case, the amount of downward motion can be computed using the magnitude of the derivative of the sequence of frames respect to the time $|\partial S/\partial t|$, 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 $\partial|\partial S/\partial t|/\partial t$. Figure 4 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 $\sum(\partial|\partial S/\partial t|/\partial t)$ 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.

Finally, the NOF feature is computed as follows:

$$\mathrm{NOF}^1 = \frac{\sum_i VM^1_i}{\max(\sum_i VM^1_i + \sum_i VM^2_i, 1)}, \qquad \mathrm{NOF}^2 = 1 - \mathrm{NOF}^1. \tag{9}$$

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

$$V^i_{\mathrm{MM}_k} := \min\left(1, V^i_{\mathrm{MM}_k}\right), \quad \forall k \in \{1, \ldots, e\}, \quad G = \frac{\left(V^i_{\mathrm{MM}} \cdot V^i_{\mathrm{GM}}\right)}{\sum_k V^i_{\mathrm{MM}_k}}, \tag{10}$$

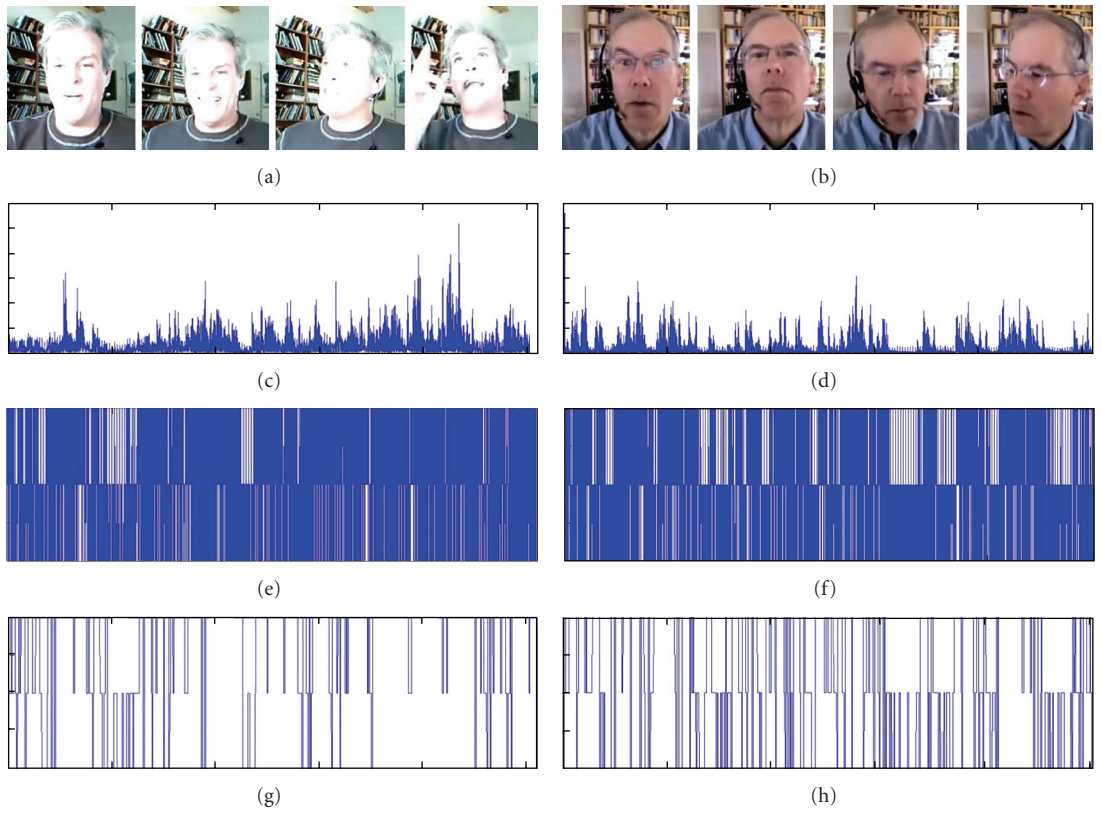

Figure 2: (a, b) Two speakers, (c, d) initial global movement, (e, f) 3-levels post-processing, and (g, h) filtering using window slicing, respectively. The x-axis corresponds to the frame number.

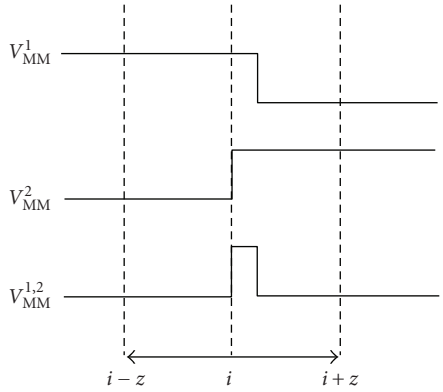

Figure 3: Interruption measurement.

where $i \in \{1,2\}$ is the speaker, $k \in \{1,\ldots,e\}$, and "$\cdot$" for the vector scalar product. 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:

$$\mathrm{SGD}^1 = \frac{\sum_i G_i^1}{\max\left(\sum_i G_i^1 + \sum_i G_i^2, 1\right)}, \qquad \mathrm{SGD}^2 = 1 - \mathrm{SGD}^1. \tag{11}$$

2.4. Interest-Based Indicators. In [19], the authors define a set of interaction-based features obtained from audio information. These features have been proved to be useful in many general social signal experiments. Thus, in this paper, we reformulate these features from a visual point of view using the movement-based features defined at the previous section.

(i) Speaking Time or Activity—ST. This features are computed for each speaker separately as described in the previous section.

(ii) Speaking Engagement—E. This feature refers to the involvement of a participant in the communication. In this case, we compute the engagement based on the activity of both speakers' mouths. Then, this feature is computed as

$$E = V_1^{\mathrm{MM}} \cdot V_2^{\mathrm{MM}}, \tag{12}$$

where "$\cdot$" stands for the scalar product between vectors, and V_1^{MM} and V_2^{MM} are the mouth movement vectors of first and second speakers, respectively.

(iii) Corporal Engagement—M. This feature refers to when one participant subconsciously *copies* another participant behavior. We approximate this feature as

$$M = V_1^{\mathrm{GM}} \cdot V_2^{\mathrm{GM}} + V_1^{\mathrm{FM}} \cdot V_2^{\mathrm{FM}} + V_1^{\mathrm{BM}} \cdot V_2^{\mathrm{BM}} \tag{13}$$

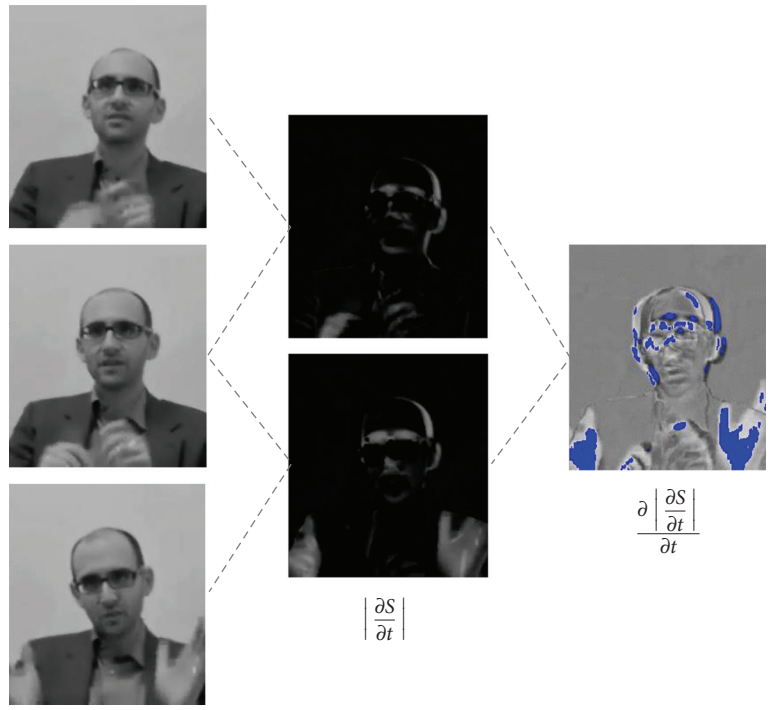

Figure 4: Vertical movement approximation.

taking into account that we consider that engagement appears when there exists simultaneous activity of face, body, or global movement, being V^{GM}, V^{FM}, and V^{BM} the global, face, and body movement vectors, respectively.

(iv) Stress—S. This feature refers to the variation in emphasis (that is, the amount of corporal movement of a participant while he is speaking). We compute this feature as

$$\forall k \in \{1,\ldots,e\}, \quad \begin{aligned} V_{i,k}^{\mathrm{MM}} &:= \min\left(1, V_{i,k}^{\mathrm{MM}}\right), \\ S &= \frac{\left(V_i^{\mathrm{MM}} \cdot V_i^{\mathrm{GM}}\right)}{\sum_k V_{i,k}^{\mathrm{MM}}}, \end{aligned} \tag{14}$$

where $i \in \{1,2\}$ is the speaker, $k \in \{1,\ldots,e\}$, and V^{GM} and V^{MM} are the global and mouth movement vectors, respectively. This measure corresponds to the global movement of each person only taking into account when he is speaking, and normalizing this value by the speaking time. This feature is computed for each speaker separately (S_1 and S_2).

3. Learning Dominance and Interest Indicators of Face-to-Face Interactions

In this paper, we define the dominance detection problem as a two-class categorization task. Although we realize that the dominance can be nonsignificative or ambiguous in some conversations, we base on those cases where there exists a clear agreement among observer's opinion when detecting the dominant people. On the other hand, in the case of the observer's interest, we define a three-level classification problem. In order to predict the degree of interest of a new observer when looking at a particular face-to-face interaction, we base on Error-Correcting Output Codes. In this section, we briefly overview the details of this framework.

3.1. Error-Correction Output Codes. The Error-Correcting Output Codes (ECOC) framework [24] is a simple but powerful framework to deal with the multiclass categorization problem based on the embedding of binary classifiers. Given a set of N_c classes, the basis of the ECOC framework consists of designing a codeword for each of the classes. These codewords encode the membership information of each binary problem for a given class. Arranging the codewords as rows of a matrix, we obtain a "coding matrix" M_c, where $M_c \in \{-1,0,1\}^{N_c \times k}$, being k the length of the codewords codifying each class. From the point of view of learning, M_c is constructed by considering k binary problems, each one corresponding to a column of the matrix M_c. Each of these binary problems (or dichotomizers) splits the set of classes in two partitions (coded by $+1$ or -1 in M_c according to their class set membership, or 0 if the class is not considered by the current binary problem).

At the decoding step, applying the k trained binary classifiers, a code is obtained for each data point in the test set. This code is compared to the base codewords of each class defined in the matrix M_c, and the data point is assigned to the class with the "closest" codeword.

Figure 5 shows the one-versus-one ECOC configuration [25, 26] for a 4-class problem. The white positions are coded to $+1$, the black positions to -1, and the grey positions

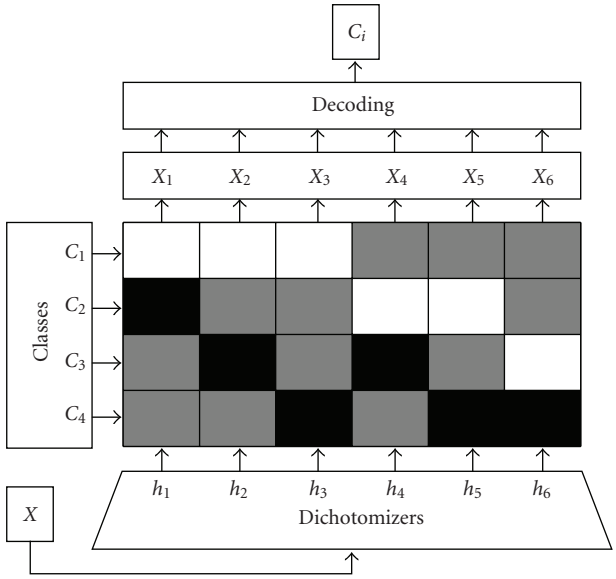

Figure 5: One-versus-one ECOC design for a 4-class problem.

correspond to the zero symbol, which means that the class is not considered by its corresponding dichotomizer. In the case of the one-versus-one design, given N_c classes, $N_c(N_c - 1)/2$ dichotomizers are trained during the coding step splitting each possible pair of classes. Then, at the decoding step, when a new test sample arrives, the previously leant binary problems are tested, and a codeword $[X_1, \ldots, X_6]$ is obtained and compared to the class codewords $\{C_1, \ldots, C_4\}$, classifying the new sample by the class C_i which codeword minimizes the decoding measure.

In our case, though different base classifiers can be applied to the ECOC designs, we use the Gentle version of Adaboost on the one-versus-one ECOC design [24]. We use Adaboost since at the same time that it learns the system splitting classes it works as a feature selection procedure. Then, we can analyze the selected features to observe the influence of each feature to rank the perceived interest of dyadic video communication. Concerning the decoding strategy, we use the Loss-weighted decoding [27], which has recently shown to outperform the rest of state-of-the-art decoding strategies.

4. Experiments and Results

In order to evaluate the performance of the proposed methodology, first we discuss the data, methods, validation protocol, and experiments.

(i) Data. The data used for the experiments consists of dyadic video sequences from the public New York Times opinion video library [22]. In each conversation, two speakers with different points of view discuss about a specific topic (i.e., "In the fight against terrorism, is an American victory in sight?"). From this data set, 18 videos have been selected. These videos are divided into two mosaics of nine videos to avoid the bias introduced by the order of visualization. The two mosaics are shown in Figure 6. To compare videos at similar conditions, all speakers are mid-age men. Each video has a frame rate of 12 FPS and a duration of four minutes, which corresponds to 2880 frames video sequences.

(ii) Methods:

(a) Dominance. In order to train a binary classifier to learn the dominance features (ST, NSI, NOF, and SGD), we have used different classifiers: Gentle Adaboost with 100 decision stumps [28], Linear Support Vector Machines with the regularization parameter $C = 1$ [29], Support Vector Machines with Radial Basis Function kernel with $C = 1$ and $\sigma = 0.5$ [29], Fisher Linear Discriminant Analysis using 99% of the principal components [30], and Nearest Mean Classifier.

(b) Interest. We compute the six interaction-based interest features ST_1, ST_2, E, S_1, S_2, and M for each of the 18 previous dyadic sequences. The one-versus-one Error-Correcting Output coding design [24] with Exponential Loss-Weighted decoding [27] and 100 runs of Gentle Adaboost [23] base classifier is used to learn the interest categories.

(iii) 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 and interest labels among observers answers, the indicators described in previous sections were automatically computed and used to learn the observer's opinion.

(iv) Validation Protocol. We apply leave-one-out and bootstrap evaluation and test for the confidence interval at 95% with a two-tailed t-test. We also use the Friedman test to look for statistical difference among observers' interest.

4.1. Observers Inquiries. We performance two inquiries, one asking for the dominant people and another one asking to rank the interest of dyadic conversations.

4.1.1. Dominance Inquiry. We performed a study with 40 people from 13 different nationalities asking for their opinion regarding the most dominant people at each New York Times dyadic conversation. The observers labeled 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 assess the reliability of agreement between the raters, we apply Kappa statistic. However, since the Kappa statistic is designed to compute the agreement between just two raters, we use the Fleiss' Kappa, a generalization of Scott's pi statistic and related to Cohen's Kappa statistic, that works

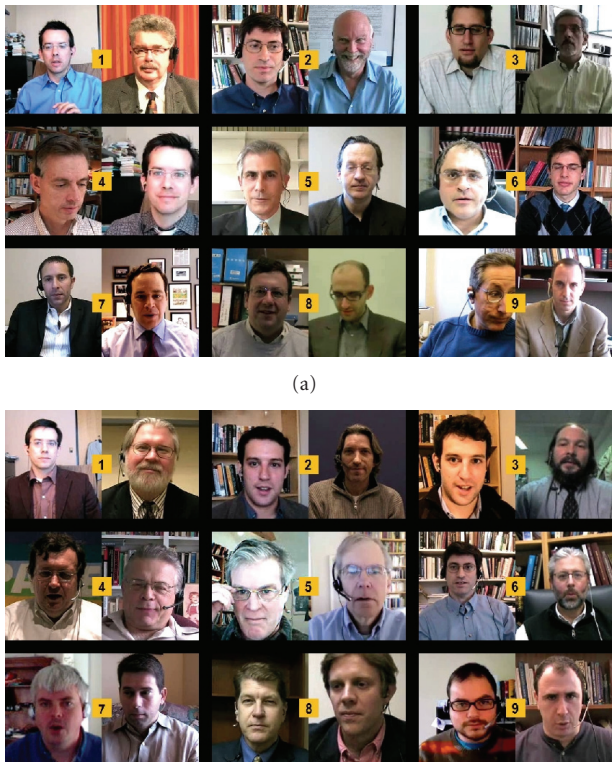

Figure 6: Mosaics of dyadic communication.

for any number of raters giving categorical ratings to a fixed number of items [31].

In our case, with 40 raters, 18 videos, and two possible categories (dominant speaker), using the rating results, we obtained a k-value of 0.55. In the six-level Fleiss' Kappa interpretation, this value corresponds near to substantial agreement.

However, it is important to make clear that dominance can be ambiguous in some situations. In fact, our initial data was composed by 20 video sequences, from which we removed the two ones with more disagreement among raters.

4.1.2. Interest Inquiry. In order to rank the interest of conversations of Figure 6, the 40 people categorized the videos of both mosaics, separately, from one (highest perceived interest) to nine (lowest perceived interest). In each mosaic, the nine conversations are displayed simultaneously during four minutes, omitting audio. The only question made to the observers was "In which order would you like to see the following videos based on the interest you feel for the conversation?" Table 1 shows the mean rank and confidence interval of each dialog considering the observers' interest. The ranks are obtained estimating each particular rank r_i^j for each observer i and each video j, and then, computing the mean rank R for each video as $R_j = (1/P)\sum_i r_i^j$, where P is the number of observers. The confidence intervals are computed with a two-tailed t-test at 95% of the confidence level.

Note that for each mosaic there exist low and high values defining different levels of expected interest. Moreover, the low magnitude of the confidence intervals also shows that there exists some "agreement" among the levels of perceived interest by the raters. These mean ranks will be used in the next experiments to perform an automatic multi-class classification of perceived interest.

4.2. Dominance Evaluation. For the dominance experiments, first, we compare the observer's opinion with a manual labeled procedure. And second, we perform an automatic dominant classification procedure.

4.2.1. Labeled Data. In order to analyze the dominance indicators defined at the previous sections, we manually annotated them for the dyadic video sequences. For each four-minute 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 (3456 manual values considering the set of eighteen 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 value is set to one for that participant and that interval, independently of its duration, otherwise it is set to zero.

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 (6), (8), (9), and (11), respectively. Some numerical results for videos of the first mosaic in Figure 6 are shown in the blue bars of Figure 7.

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 2. 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 most of the videos as expected by the observers.

4.2.2. Automatic Dominance Detection. For this experiment, we automatically computed the ST, NSI, NOF, and SGD dominance indicators as explained in the previous section. The videos are in 12 FPS, and four minutes per video defines independent sequences of 2880 frames, representing a total of 51840 analyzed frames. The mouth history in frames and

Table 1: Ranking positions and confidence interval of dyadic interactions.

	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9
Mosaic 1	5.4 (1.0)	5.3 (0.8)	4.3 (0.9)	3.3 (0.6)	2.7 (0.6)	6.7 (0.8)	6.4 (1.0)	3.1 (1.0)	7.9 (0.6)
Mosaic 2	3.4 (0.9)	4.3 (0.8)	4.8 (0.9)	7.2 (1.0)	4.2 (1.2)	5.9 (1.0)	4.2 (1.0)	6.8 (0.8)	4.3 (0.9)

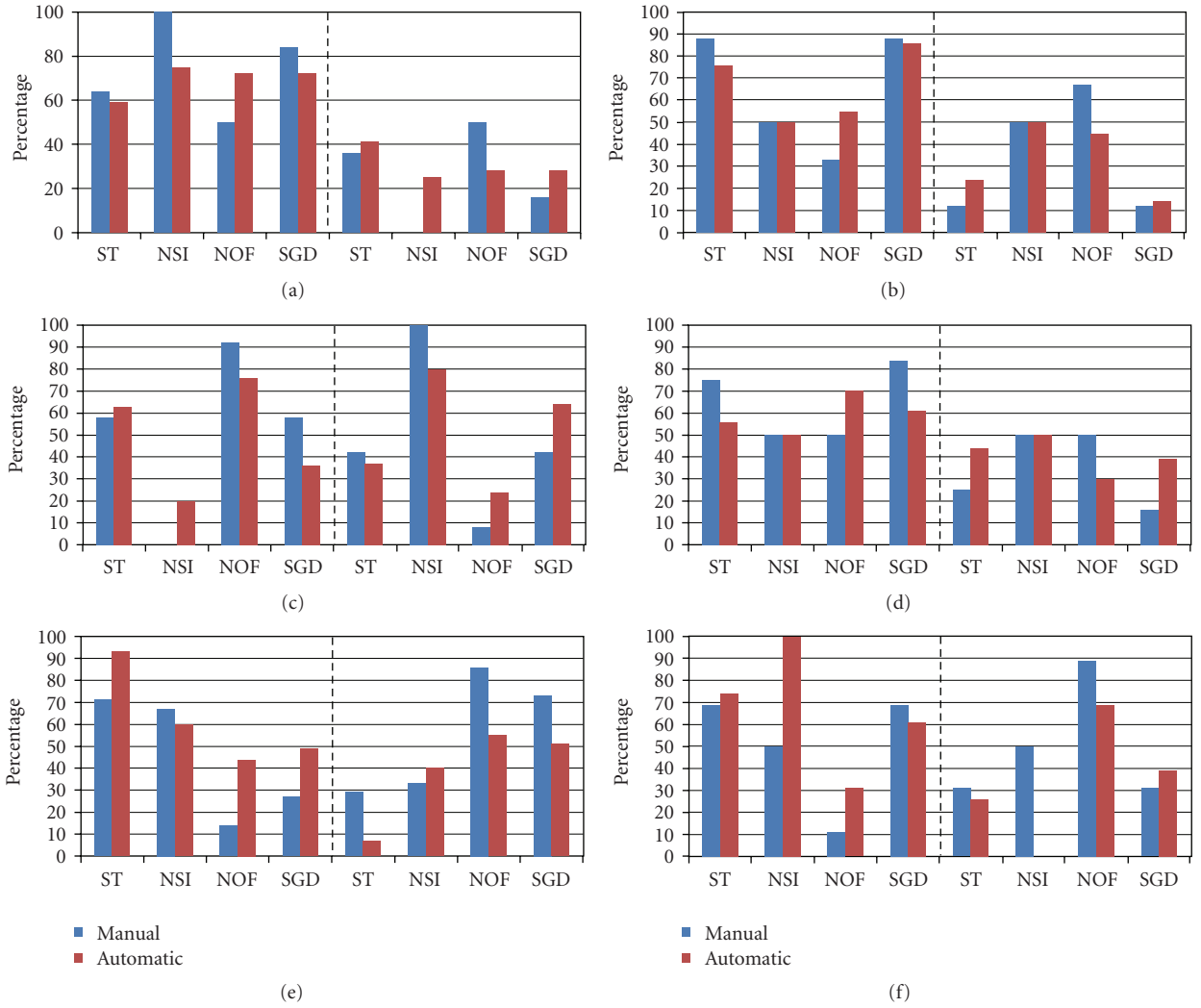

Figure 7: Manual (blue) and automatic (red) dominance indicators values.

Table 2: Dominance classification results using independent manually-labeled indicators.

Indicator	Accuracy
Manual ST	96%
Manual NSI	83%
Manual NOF	74%
Manual SGD	74%

the windows size for the successful interruption computation are set to ten. Some numerical obtained values are shown in the red bars of Figure 7 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 in 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 3. 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,

Table 3: Dominance classification results using independent automatic-extracted dominance indicators.

Indicator	Accuracy
Automatic ST	96%
Automatic NSI	78%
Automatic NOF	74%
Automatic SGD	74%

Table 4: Dominance classification results using dominance indicators and leave-one-out evaluation (first column) and bootstrap evaluation (second column).

Learning strategy	Accuracy	Accuracy
Gentle Adaboost	100%	91.8%
Linear SVM	82.3%	86.8%
RBF SVM	100%	85.9%
FLDA	100%	91.8%
NMC	82.3%	75.7%

we used a set of classifiers, performing two experiments. The first experiment corresponds to a leave-one-out evaluation, and the second one to a bootstrap [32] evaluation. To perform a bootstrap evaluation, 200 random sequences of videos were defined, where each sequence has 18 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 eighteen performances for the eighteen 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 4. 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.

4.3. Interest Evaluation. For the interest quantification problem, we define a 3-class problem based on the results obtained from the observer's interest opinion rank.

4.3.1. Automatic Ranking of Interest of Dyadic Sequences. After computing the mean rank obtained by observers' rating, we define a multi-class categorization problem for each of the two mosaics. In each case, three categories are determined using the observers' ranks: high, medium, and low interest. The categories are shown in Table 5. For each

Table 5: Interest categories for the two mosaics of Figure 6 based on the observers' criterion.

	High interest	Medium interest	Low interest
M.1	5–2.7 (0.6)	3–4.3 (0.9)	7–6.4 (1.0)
	8–3.1 (1.0)	2–5.3 (0.8)	6–6.7 (0.8)
	4–3.3 (0.6)	1–5.4 (1.0)	9–7.9 (0.6)
M.2	1–3.4 (0.9)	9–4.3 (0.9)	6–5.9 (1.0)
	5–4.2 (1.2)	2–4.3 (0.8)	8–6.8 (0.8)
	7–4.2 (1.0)	3–4.8 (0.9)	4–7.2 (1.0)

mosaic, the number of the videos with its corresponding mean rank and confidence interval is shown. One can see that in the case of the first mosaic there exist three clear clusters, meanwhile in the case of the second mosaic, though the low interest category seems to be split from two first categories, high and medium categories are not clearly discriminable in terms of their mean ranks.

Now, we use the one-versus-one ECOC design with Exponential Loss-weighted decoding to test the multi-class system. For each mosaic, we used eight samples to learn and the remaining one to test, and repeat for each possibility (nine classifications). For each sequence, the six interaction-based interest features A_1, A_2, E, S_1, S_2, and M are computed based on the movement-based features. Concerning the movement-base features, the values are computed among consecutive frames, and the faces are detected using a cascade of weak classifiers of six levels with 100 runs of Gentle Adaboost with decision stumps, considering the whole set of Haar-like features computed on the integral image. 500 positive faces were learnt against 3000 negative faces from random Google background images at each level of the cascade. Finally, the size of the windows for the post-processing of movement-based vectors was $q = 5$. The obtained results are shown in the following confusion matrices:

$$\mathrm{CM}_1 = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 3 \end{pmatrix}, \qquad \mathrm{CM}_2 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{15}$$

for the two mosaics, respectively. In the case of the first mosaic, six from the nine video samples were successfully classified to their corresponding interest class. In the case of the second mosaic, five from the nine categories were correctly categorized. These percentages show that the interaction-based features are useful to generalize the observers' opinion.

Furthermore, missclassifications involving adjacent classes can be partially admissible. Note that nearer classes have nearer interest rank than distant classes. In order to take into account this information, we use the distances among neighbor classes centroids to measure an error cost EC: $\mathrm{EC}(C_i, C_j) = d_{ij} / \sum_k d_{ik}$, where EC estimates the error cost of classifying a sample from class C_i as class C_j. The term d_{ij} refers to the Euclidean distance between centroids of classes C_i and C_j, and $k \in [1, 2, 3] \setminus i$ in the case of three categories. Note that this measure returns a value of zero if the decision is true and an error cost relative to the distance to the correct

class C_j, being one if the predicted class is not adjacent to the correct one. Then, applying the previous measure to our two 3-class problems, we obtain the following error cost matrices:

$$\mathrm{EC}_{\mathrm{CM}_1} = \begin{pmatrix} 0 & 0.49 & 1 \\ 0.49 & 0 & 0.51 \\ 1 & 0.51 & 0 \end{pmatrix}, \quad \mathrm{EC}_{\mathrm{CM}_2} = \begin{pmatrix} 0 & 0.2 & 1 \\ 0.2 & 0 & 0.8 \\ 1 & 0.8 & 0 \end{pmatrix}. \tag{16}$$

If we use the information from the previous confusion matrices and the error cost matrices, we can estimate a *relative* performance RF for the first mosaic of RF = 83.38% and of RF = 82.30% for the second mosaic. Moreover, in 17 of the 18 dyadic sequences analyzed, features related to the mouth and body movement are selected by the Adaboost ECOC base classifier. In particular, the stress feature seems to maximize the agreement among the observers' ranks. Thus, it shows to be one of the most important features to obtain a correct interest rank, as expected.

5. Conclusions

We analyzed a set of non-verbal cues to detect the dominant people and the level of interest from the point of view of observers in face-to-face video sequences. We performed an experiment with 40 observers asking for their opinion regarding the most influent participant and interest of a set of dyadic sequences. Results showed high agreement among observers opinion. We also defined a set of gestural communication indicators and manually annotated the videos. Comparing to the observers opinion, the indicators have shown high discriminative power. Moreover, an automatic approximation to the dominant and interest indicators based on low-level movement-based features was presented. Adaboost and the Error-Correcting Output Codes framework were used to detect the dominant people and learn to rank the perceived interest of face-to-face interactions. The automatic system has shown a good correlation between the automatic categorization results and the manual labeling. In particular, the learning system showed that stress features have a high predictive power for ranking observer's interest meanwhile the speaking time is the preferred one to detect dominant people when looking at face-to-face interactions.

The simple set of considered features obtained high performance capturing the motion information in the analyzed video sequences. However, as future work we plan to extend our framework to changing environments where the noncontrolled conditions will require more complex motion-based feature-extraction methodologies.

Acknowledgment

This work has been supported in part by projects TIN2009-14404-C02 and CONSOLIDER-INGENIO CSD 2007-00018.

Endnotes

1. One of the most immediate uses of this definition of interest appears in the context of TV or other mass-media providers in measuring audience interest in their content (i.e., political debates, interviews, etc.). Moreover, this can also be used in content retrieval; in particular, the measure of perceived interest can be used to sort out or rank video-lectures, speeches, or debates retrieved via web.

References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[3] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, "An emotion-aware voice portal," in *Proceeding of the Electronic Speech Signal Processing (ESSP '05)*, Prague, Czech Republic, 2005.

[4] P. Ekman, *Emotions Revealed*, Times Books, New York, NY, USA, 2003.

[5] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI '08)*, pp. 53–60, October 2008.

[6] B. Lepri, N. Mana, A. Cappelletti, and F. Pianesi, "Automatic prediction of individual performance from "thin slices" of social behavior," in *Proceedings of International Conference on Multimedia (ACM '09)*, pp. 733–736, New York, NY, USA, October 2009.

[7] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro, "Modeling the personality of participants during group interactions," in *Proceedings of International Conference on User Modeling, Adaptation, and Personalization (UMAP '09)*, vol. 5535, pp. 114–125, Trento, Italy, June 2009.

[8] K. Truong and D. A. van Leeuwen, "Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features," in *Proceedings of Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, August 2007.

[9] D. Gatica-Perez, "Analyzing group interactions in conversations: a review," in *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 41–46, 2006.

[10] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, 2005.

[11] S. L. Ellyson and J. F. Dovidio, *Power, Dominance, and Nonverbal Behavior*, Springer, Berlin, Germany, 1985.

[12] R. Rienks and D. Heylen, "Automatic dominance detection in meetings using svm," in *Proceedings of Machine Learning for Multimodal Interaction (MLMI '05)*, Edinburgh, UK, May 2005.

[13] D. Babu, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[14] H. Hung, D. Babu, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Investigating automatic dominance estimation in groups from visual attention and speaking activity," in *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI'08)*, pp. 233–236, Chania, Greece, 2008.

[15] H. Hung, D. Jayagopi, C. Yeo, et al., "Using audio and video features to classify the most dominant person in a group meeting," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 835–838, 2007.

[16] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: a review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[17] D. Gatica-Perez, "Analyzing group interactions in conversations: a review," in *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI '06)*, pp. 41–46, September 2006.

[18] J. S. Efran, "Looking for approval: effects on visual behavior of approbation from persons differing in importance," *Journal of Personality and Social Psychology*, vol. 10, no. 1, pp. 21–25, 1968.

[19] A. Pentland, "Socially aware computation and communication," *Computer*, vol. 38, no. 3, pp. 33–40, 2005.

[20] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behaviour: a survey," in *Proceedings of International Conference on Multimodal Interfaces (ICMI '06)*, pp. 239–248, Alberta, Canada, November 2006.

[21] D. Gatica-Perez, L. McCowan, S. B. D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 1, pp. 489–492, Philadelphia, Pa, USA, March 2005.

[22] http://video.nytimes.com/.

[23] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[24] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via errorcorrecting output codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–282, 1995.

[25] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: a unifying approach for margin classifiers," *The Journal of Machine Learning Research*, vol. 1, no. 2, pp. 113–141, 2001.

[26] T. Hastie and R. Tibshirani, "Classification by pairwise grouping," in *Proceedings of Neural Information Processing Systems (NIPS '98)*, vol. 26, pp. 451–471, Denver, Colo, USA, November 1998.

[27] S. Escalera, O. Pujol, and P. Radeva, "On the decoding process in ternary errorcorrecting output codes," *Transactions in Pattern Analysis and Machine I Intelligence*, vol. 32, no. 1, pp. 120–134, 2010.

[28] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[29] Osu-svm-toolbox, http://svm.sourceforge.net/docs/3.00/api/.

[30] P. Tool, http://prtools.org/.

[31] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[32] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, Chapman & Hall, Boca Raton, Fla, USA, 1993.