Measuring *Interest* of Human Dyadic Interactions

Sergio ESCALERA, Oriol PUJOL, Petia RADEVA, and Jordi VITRIÀ
Dept. Matemàtica Aplicada i Anàlisi, Gran Via 585, 08007, Barcelona
Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Barcelona
{sergio,oriol,petia,jordi}@maia.ub.es

Abstract. In this paper, we argue that only using behavioural motion information, we are able to predict the *interest* of observers when looking at face-to-face interactions. We propose a set of movement-related features from body, face, and mouth activity in order to define a set of higher level interaction features, such as stress, activity, speaking engagement, and corporal engagement. Error-Correcting Output Codes framework with an Adaboost base classifier is used to learn to rank the perceived observer's *interest* in face-to-face interactions. The automatic system shows good correlation between the automatic categorization results and the manual ranking made by the observers. In particular, the learning system shows that stress features have a high predictive power for ranking *interest* of observers when looking at of face-to-face interactions.

1. Introduction

For a long time, the scientific community has been focused on computer vision and speech domains for analyzing individuals as isolated items. However, humans are social beings by nature. Social communication relies on two communication channels. The direct channel is related to the conscious level of the emitter and deals with semantic conventions; Indirect channels are unconscious cognitive processes based mostly on nonverbal communication cues – facial expressions, hand gestures, body postures, dynamic speech patterns. For most of us, this 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, establishing credibility, or leadership. Thus, understanding these social signals is the basis for understanding human-to-human interaction. Artificial social perception is the discipline devoted to the analysis of these signals and the social messages they convey. Nowadays, works in artificial social perception are grouped according to their underlying social theory in either emotion-based signaling or human social signaling. The first trend [4] considers that affective states are a different source of messages communicated by social signals. Following this line of thought different works have been done considering emotions, such as emotion description, emotion detection, or emotion-based applications [1].

On the other hand, human social signaling considers a wider set of cues as facilitators of the communication between people. These 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 non-linguistic vocalizations like laughs, cries, sighs, and coughs [12]. However, humans group these basic signals to form social messages (i.e. dominance, trustworthiness, friendliness, etc). The detection of social messages has recently received attention – i.e. in [8] *dominance* is estimated from multimodal sources.

In this article, we give an approximation of the quantification of *interest* from the point of view of an external observer exclusively analyzing visual cues. The term *interest* 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 [6]. 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.

With the purpose of quantifying the level of *interest* of an external observer, we base our features in the works of Pentland et al.[11,10]. In their research, the authors propose a small set of social signals, such as activity level, stress, speaking engagement, and corporal engagement for analyzing nonverbal *speech* patterns during dyadic interactions. We extend and propose an implementation of the concepts proposed in those works. We identify different basic social signals – motion-related features from body, face, and mouth activity – that allow to build up the four high level interaction features – stress, activity, speaking engagement, and corporal engagement. We argue that only using behavioural motion information, we are able to predict the perceived *interest* by observers. These features are included in an Error-Correcting Output Codes design, which learn the different levels of *interest* perceived by observers when looking at face-to-face interactions. The automatic system shows good correlation between the automatic categorization results and the manual ranking made by the observers.

The layout of the article is as follows: Section 2 describes the basic social signal features extracted from the videos. We propose an approach to compute the higher level interaction social signals starting from those basic features. Section 3, introduces the machine learning framework used in the paper. Section 4 shows the experimental settings and results. Finally, section 5 concludes the paper.

2. Visual dyadic features

In order to predict the level of *interest* perceived by observers when looking at face-to-face interaction video sequences, first, we 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 movement features. These features will serve as bases to build higher level interaction features, namely stress, activity, speaking engagement, and corporal engagement.

2.1. Movement-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 four individual signal features: global movement, face movement, body movement, and mouth movement.

• **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:

$$GM_{ij} = \sum_k |s_{j,k} - s_{i,k}| \tag{1}$$

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

• **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 & Jones [9] 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 [9]. The Haar-like features and the rotated ones have been used to define the feature space [9]. Figure 1(c) shows an example of a detected face of size $n \times m$, in the ith frame of a sequence, denoted by $F_i \in \{0, .., 255\}^{n \times m}$. Then, the face movement feature FM_{ij} at ith frame is defined as follows:

$$FM_{ij} = \frac{1}{n \cdot m} \sum_k |F_{j,k} - F_{i,k}| \tag{2}$$

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

• **Body movement**: We define the body movement BM as follows:

$$BM_{ij} = \sum_k |s_{i,k} - s_{j,k}| - \sum_{f_k \in F^{ij}} f_k \tag{3}$$

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 Fig. 1(e).

• **Mouth movement**: In order to avoid the bias that can appear due to the translation of mouth detection between consecutive frames, 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 $F_i \in \{0, .., 255\}^{n \times m}$ 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 movement feature MM_{iL} is computed as follows:

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

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. 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 Fig. 1(f), and its corresponding accumulated substraction for $L = 3$ is shown in Fig. 1(g).

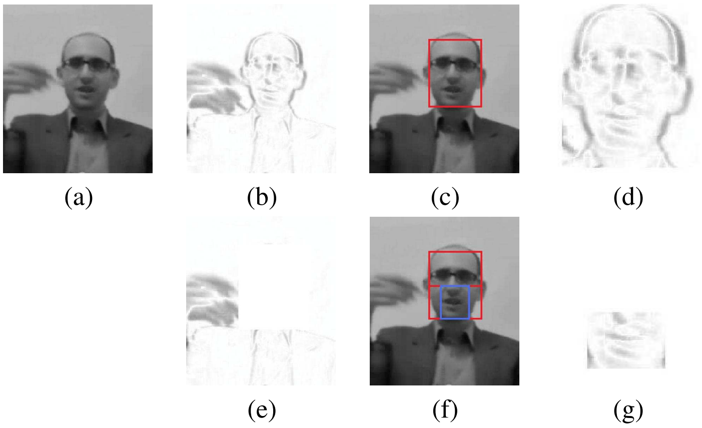


Figure 1. (a) ith frame from dialog, (b) Global movement GM_{ij}, (c) Detected face F_i, (d) Face movement FM_{ij}, (e) Body movement BM_{ij}, (f) Mouth detection M_i, and (g) Mouth movement MM_{ij}.

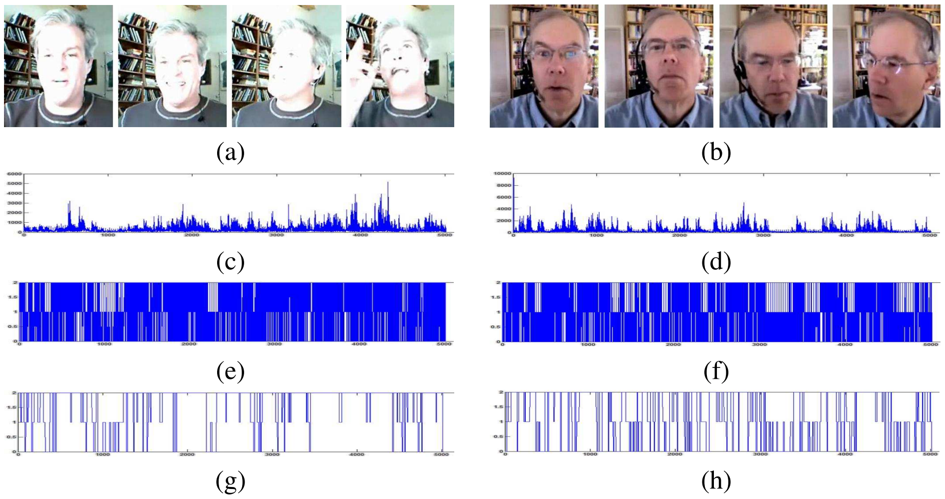
2.2. Post-processing

After computing the values of GM_{ij}, FM_{ij}, BM_{ij}, and MM_{iL} for a sequence of e frames ($i, j \in [1, .., e]$), we filter the responses. Fig. 2(c) and (d) correspond to the global movement features GM_{ij} in a sequence of 5000 frames at 12 FPS for the speakers of Fig. 2(a) and (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:

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

The result of this step is shown in Fig. 2(e) and (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 (Fig. 2(g) and (h), respectively).



(a) (b) (c) (d) (e) (f) (g) (h)

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.

2.3. Interaction-based indirect features

In [11], the authors define a set of interaction-based features obtained from audio information. In this paper, we re-formulate these features from a visual point of view using the movement-based features defined at the previous section.

• **Activity**: This feature refers to how much an emitter participates in the conversation. We compute this feature as:

$$A = \sum_{k} V_i^{MM}, k \in \{1, .., e\} \tag{6}$$

where V_i^{MM} is the mouth movement vector of speaker $i \in \{1, 2\}$. This measure corresponds to the total mouth movement, codifying the speaking time weighted by the movement degree. This feature is computed for each speaker separately (A_1 and A_2).

• **Speaking engagement**: 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:

$$E = V_1^{MM} \cdot V_2^{MM} \tag{7}$$

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 speaker, respectively.

• **Corporal engagement**: This feature refers to when one participant subconsciously *copies* another participant behavior. We approximate this feature as:

$$M = V_1^{GM} \cdot V_2^{GM} + V_1^{FM} \cdot V_2^{FM} + V_1^{BM} \cdot V_2^{BM} \tag{8}$$

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.

• **Stress**: 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:

$$\forall k \in \{1,..,e\}, V_{i,k}^{MM} := \min(1, V_{i,k}^{MM}), S = \left(V_i^{MM} \cdot V_i^{GM}\right) / \sum_k V_{i,k}^{MM} \quad (9)$$

where $i \in \{1,2\}$ is the speaker, $k \in \{1,..,e\}$, 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 to rank the *interest* of face-to-face interactions

In this paper, we split the observer's *interest* in three levels. In order to predict the degree of *interest* of a new observer when looking at a particular face-to-face interaction, we define a multi-class categorization procedure based 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 [3] is a simple but powerful framework to deal with the multi-class 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.

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 [3]. 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 [5], 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.

• $Data$: The data used for the experiments consists of dyadic video sequences from the public New York Times opinion video library [7]. 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 Fig. 3. 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 seven minutes, which corresponds to 5040 frames video sequences.

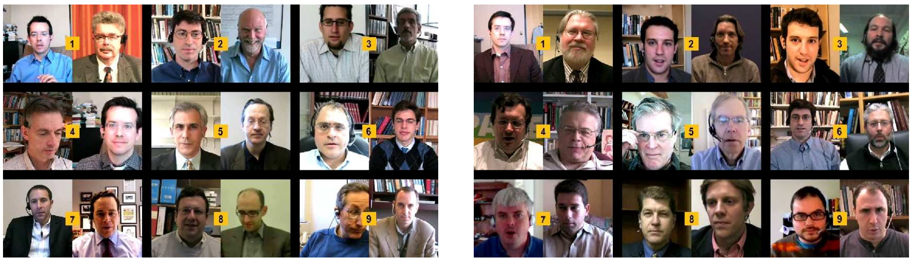


Figure 3. Mosaics of dyadic communication.

• $Methods$: We compute the six interaction-based indirect features A_1, A_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 [3] with Exponential Loss-Weighted decoding [5] and 100 runs of Gentle Adaboost [9] base classifier is used to learn the *interest* categories.

• *Validation protocol*: We apply two 9-fold cross-validation (one for each mosaic of 9 videos) 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*.

• *Experiments*: First, we analyze the correlation among observers when ranking the *interest* in both mosaics scenarios. And second, we perform an automatic ranking using the interaction-based features based on the observers' decisions.

4.1. *Analyzing observers' ranking*

In order to rank the conversations of Fig. 3, 40 people from 10 different nationalities categorized the videos of both mosaics, separately, from one (highest *interest*) to 9 (lowest *interest*). In each mosaic, the nine conversations are displayed simultaneously during seven 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 think the conversation has?" 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 $R_j = \frac{1}{P}\sum_i r_i^j$, where P is the number of observers.

	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)

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

In order to reject the null hypothesis that the measured ranks are due to randomness in the results, we use the Friedman test. The Friedman statistics value is computed as:

$$X_F^2 = \frac{12P}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{10}$$

In our case, with $k = 9$ videos to compare on each mosaic, $X_F^2 = 136.8$ for the first mosaic and $X_F^2 = 78.7$ for the second mosaic. Since this value is undesirable conservative, Iman and Davenport proposed a corrected statistics [2]:

$$F_F = \frac{(P-1)X_F^2}{P(k-1) - X_F^2} \tag{11}$$

Applying this correction, we obtain $F_F = 29.1$ and $F_F = 12.7$ for the two mosaics, respectively. With nine videos and 40 observers, F_F is distributed according to the F distribution with 8 and 312 degrees of freedom. The critical value of $F(8, 312)$ for 0.05 is 2.18. As the value of F_F for both mosaics is higher than 2.18, we can reject the null hypothesis. Then, we can state that there exist correlation among observers' opinion.

4.2. Automatic ranking of dyadic sequences

After determining that there exist statistical evidences confirming the correlation among observers' perceived *interest*, we define two experiments, one for each mosaic. In each case, three categories are determined using the observers' ranks: high, medium, and low *interest*. The categories are shown in Table 2. For each mosaic, the number of the video 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.

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

Table 2. *Interest* categories for the two mosaics of Fig. 3 based on the observers' criterion.

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 indirect 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:

$$CM_1 = \begin{pmatrix} 2\,1\,0 \\ 1\,1\,1 \\ 0\,0\,3 \end{pmatrix} \qquad CM_2 = \begin{pmatrix} 1\,1\,1 \\ 2\,1\,0 \\ 0\,0\,3 \end{pmatrix} \tag{12}$$

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, miss-classifications involving adjacent classes can be 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: $EC(C_i, C_j) = \frac{d_{ij}}{\sum_k d_{ik}}$, where EC estimates the error cost of classifying a sample from class C_j as class C_i. The term d_{ij} refers to the Euclidean distance between centroids of classes C_i and C_j, and $k \in [1, 2, 3]\backslash i$ 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:

$$EC_{CM_1} = \begin{pmatrix} 0 & 0.49 & 1 \\ 0.49 & 0 & 0.51 \\ 1 & 0.51 & 0 \end{pmatrix} \qquad EC_{CM_2} = \begin{pmatrix} 0 & 0.2 & 1 \\ 0.2 & 0 & 0.8 \\ 1 & 0.8 & 0 \end{pmatrix} \tag{13}$$

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 correlation 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

In this paper, we showed the correlation of observers' *interest* when they rank dyadic conversation based on indirect visual communication channels. We gave a first evidence to be able to automatically quantify observers' *interest*. We defined a set of simple motion features from body, face, and mouth activity to define a set of interaction-related features which were used to learn a set of *interest* categories from dyadic videos. Up to our knowledge, for first time ranking of subjective *interest* is automatically predicted by Computer Vision and Machine Learning techniques. Error-Correcting Output Codes framework with an Adaboost base classifier was used to learn to rank the perceived *interest* of face-to-face interactions. The automatic system shown good correlation between the automatic categorization results and the manual ranking made by the observers. In particular, the learning system showed that stress features have a high predictive power for ranking observer's *interest* when looking at face-to-face interactions, being complemented by engagement and activity for a more robust analysis.

6. Acknowledgements

This work has been partially supported by the projects TIN2006-15694-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

[1] F. Burkhardt, M. van Ballegooy, and R. Englert. An emotion-aware voice portal. In *ESSP*, 2005.

[2] J. Demsar. Statistical comparisons of classifiers over multiple data sets. In *Journal of Machine Learning Research*, volume 7, pages 1–30, 2006.

[3] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. In *Journal of Artificial Intelligence Research*, volume 2, pages 263–282, 1995.

[4] P. Ekman. Emotions revealed. In *Times Books*, 2003.

[5] S. Escalera, O. Pujol, and P. Radeva. On the decoding process in ternary error-correcting output codes. In *Transactions in Pattern Analysis and Machine I Intelligence*, in press.

[6] D. Gatica-Perez. Analyzing group interaction in conversations: a review. In *in Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2006.

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

[8] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2197–2200, 2008.

[9] M. Jones and P. Viola. Robust real-time face detection. In *International Journal of Computer Vision*, volume 57, pages 137–154, 2004.

[10] M.Pantic, A.Pentland, A.Nijholt, and T.Huang. Human computing and machine understanding of human behaviour: A survey. In *ACM ICMIÕ06*, pages 239–248, 2006.

[11] A. Pentland. Socially aware comp. and communication. In *Computer*, volume 38, pages 33–40, 2005.

[12] K. Truong and D. A. van Leeuwen. Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features. In *Workshop on the Phonetics of Laughter*, 2007.