Escritos de Psicología, Vol. 3, nº 2, pp. 41-45

www.escritosdepsicologia.es
ISSN 1989-3809

Automatic dominance detection in dyadic conversations

Detección automática de la dominancia en conversaciones diádicas

Sergio Escalera, Rosa M. Martínez, Jordi Vitrià, Petia Radeva, M. Teresa Anguera

Universidad de Barcelona

Disponible online 30 de abril de 2010

Dominance is referred to the level of influence that 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 the dominance detection of 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, these indicators are automatically extracted from video sequences and learnt by using binary classifiers. Results from the three analyses showed a high correlation and allows the categorization of dominant people in public discussion video sequences.

Keywords: Dominance detection; Non-verbal communication; Visual features.

La dominancia está relacionada con el nivel de influencia que una persona tiene en una conversación. El estudio de la dominancia es de especial interés en la psicología social, pero el problema de su estimación automática es un tema muy reciente en los contextos de computación social e inalámbrica. En este trabajo nos centramos en la detección de dominancia a partir del análisis automático de características visuales. Hacemos una estimación de la correlación entre los observadores al categorizar las personas dominantes en un conjunto de conversaciones cara a cara. Definimos diferentes indicadores de dominancia a partir de información gestual, los cuales también son anotados manualmente y comparados con la opinión de los observadores. Además, los indicadores considerados son extraídos de forma automática de las secuencias de vídeo y aprendidos mediante clasificadores binarios. Los resultados de los tres análisis muestran un alto grado de correlación y permiten categorizar de forma automática las personas dominantes en vídeos públicos de debates.

Palabras clave: Detección de dominancia; Comunicación no verbal; Características visuales.

Correspondencia: Sergio Escalera. Facultat de Matemàtiques, Gran Via de les Corts Catalanes, 585. 08007- Barcelona. E-mail: sergio@maia.ub.es

Dominance (Gatica-Pérez, 2006; McCowan et al., 2005) 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 (Ellyson & Dovidio, 1985), the problem of its automatic estimation is a very recent topic in the context of social and wearable computing (Babu, Hung, Yeo & Gatica-Pérez, 2008; Hung, Babu, Ba, Odobez & Gatica-Pérez, 2008; Hung et al., 2007; Rienks & Heylen, 2005). 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" (Gatica-Pérez, in press).

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 computing (Efran, 1968; Hung et al., 2007, 2008; Rienks & Heylen, 2005). 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 (http://video.nytimes.com). The conversations were shown to several observers that labeled the dominance based on their personal opinion. Different indicators 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 analyses –observers' opinion, manually annotated indicators, and automatic feature extraction and classification-, showed statistically significant correlation discriminating among dominant and dominated people.

First, the visual cues for dominance detection are presented.

Dominance indicators

In order to detect the dominant person in a face-to-face interaction video sequence, a set of basic visual features should be first defined.

Motion-based basic features

Given a video sequence, 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.

In order to detect the face we use the Viola & Jones face detector, and compute the face motion feature within the face region as in the case of the global motion, normalizing by the face size.

From the face region detected at frame i, the mouth region is defined in the center bottom half region of the face. Then, given the parameter l, the mouth motion feature MM_{il} is computed as the accumulative subtraction of l mouths at frames previous to i.

Post-processing

After computing the values of GM_{ij}, FM_{ij}, and MM_{il} for a sequence of e frames, 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, corresponding to low, medium, and high motion quantifications. 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}).

Dominance-based features

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.

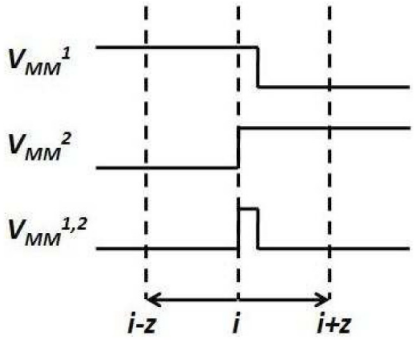
The speaker gesticulation degree - SGD: Some studies suggest that high degree of gesticulation of a participant when speaking makes the rest of participants to focus on him, being a possible indicator of dominance (also known as stress) (Pentland, 2005).

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 using the vectors computed at the previous section.

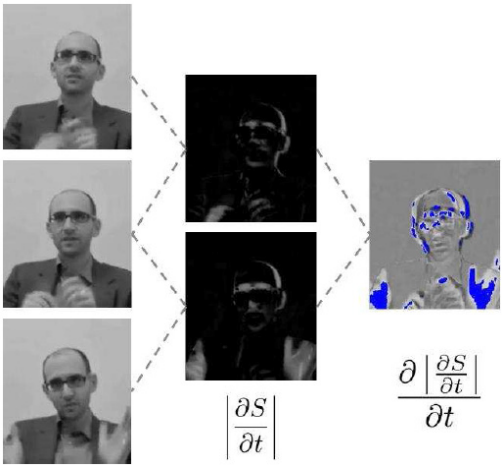
Given the 3-value mouth motion vectors $V_{MM}{}^1$ and $V_{MM}{}^1$ for both participants, we define a successful interruption I^2 of the second participant if the vector $V_{MM}{}^1$ decreases its magnitude meanwhile $V_{MM}{}^2$ increases, considering z frames for a valid interruption. An example of a successful interruption I^2 of the second speaker is shown in Figure 1.

Figure 1. Interruption measurement.



We approximate the number of times the floor is grabbed by a participant (NOF) as the amount downward motion executed by that participant. This feature can be approximated by the magnitude of the derivative of the sequence of frames respect to the time. In order to obtain the vertical movement orientation to approximate the NOF feature, we compute the derivative in time of the previous measurement. 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 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.



The speaker gesticulation degree SGD refers to the variation in emphasis. We compute this feature as the combination of face and global quantifications only taking into account the time when the participant is speaking.

For all previous indicators, the final values are then converted to percentage in order to have the measures comparable among all conversations.

Method

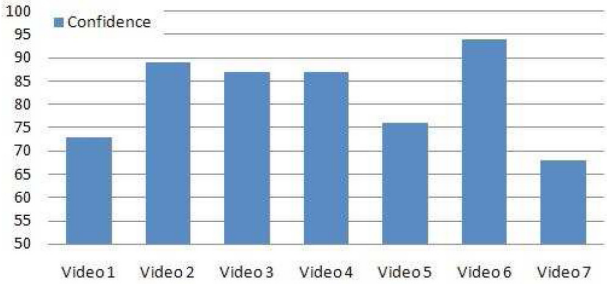
The data used for the experiments consists of dyadic video sequences from the public New York Times web site video library (http://video.nytimes.com). In each conversation, two speakers with different points of view discuss about a direct question. From this data set, seven videos have been selected. These videos are shown in Figure 3. 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.

We have used different classifiers: Discrete Adaboost with decision stumps (Fiedman, Hastie & Tibshirani, 1998), Linear SVM with the regularization parameter $C=1$ (Osu-svm-toolbox, http://svm.sourceforge.net), SVM with RBF kernel with $C=1$ and $\sigma=0.5$ (Osu-svm-toolbox, http://svm.sourceforge.net), FLDA using 99% of the principal components (http://prtools.org), and NMC.

First, we asked 40 independent observers to put a label on each of the videos. After looking for the correlation of dominance labels among the 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.

Figure 4. Observers correlation values.



Result

Observers inquiry

We performed an experiment with 40 people from 13 different nationalities asking for their opinion regarding the most dominant people. The observers labeled each dominant people for each conversation, only taking into account the visual information. The correlation results among observers opinion are shown in Figure 4. Note that all results are in the range [65,..,95] of confidence, which corresponds to high correlation among observers opinion.

Labeled data

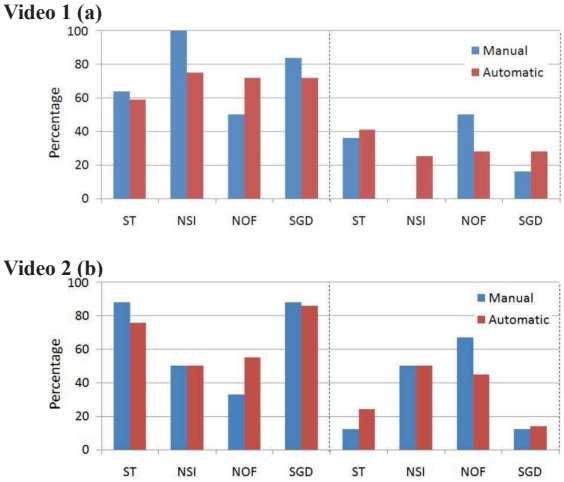
We manually annotated the indicators over the dyadic video sequences. For each four minutes video sequence, intervals of ten seconds are defined for each participant. If an indicator appears within an interval of ten seconds, the indicator is activated for that participant and that interval independently of the time the indicator appears.

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 activates the indicator, or zero otherwise. After the manual labeling, the indicators are computed by summing the manual values and estimating its percentage. The results are shown in the blue bars of Figures 5 and 6. 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.

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



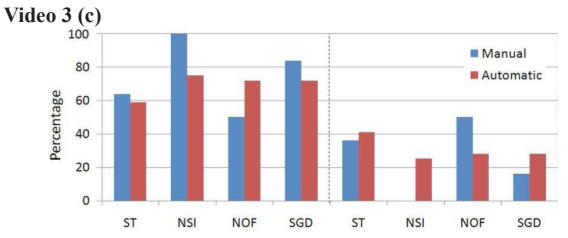


Figure 6. Manual (blue) and automatic (red) indicators values.

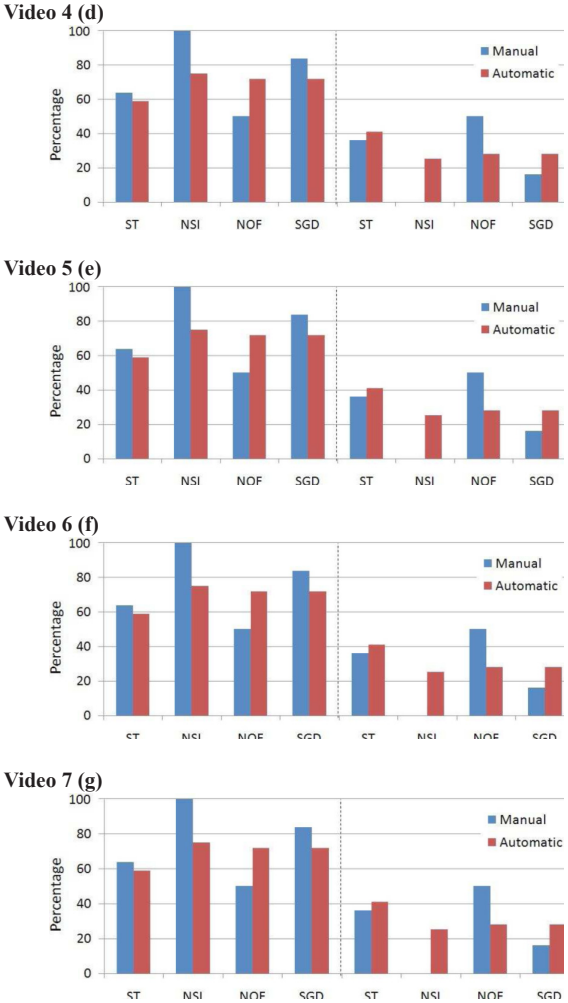


Table 1. Dominance classification results using independent manually-labeled indicators.

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

Automatic dominance features

For this experiment, we automatically computed the dominance indicators. The values obtained are shown in the red bars of Figures 5 and 6. Note that the obtained results are very simi-

lar 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 the performance results are almost the same than with manual labeling.

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

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

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 evaluation, and the second one to a bootstrap (Efron & Tibshirani, 1993) 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. The classification results are shown in Table 3. Note that all classifiers obtain results near or over 80%.

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

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 %

Discussion

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. We also defined a set of gestural communication indicators and manually annotated the videos. 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.

References

Babu, D., Hung, H., Yeo, C., & Gatica-Perez, D. (2008). Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing, 17*, 501–513.

Efran, J. S. (1968). Looking for approval: effects of visual behavior of approbation from persons differing in importance. *Journal of Personality and Social Psychology, 10*, 21–25.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Ellyson, S. L., & Dovidio, J. F. (1985). *Power, dominance, and nonverbal behavior*. New York: Springer-Verlag.

Friedman, J., Hastie, T., & Tibshirani, R. (1998). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics, 38*, 337–374.

Gatica-Perez, D. (2006). Analyzing group interactions in conversations: A review. *Multisensor Fusion and Integration for Intelligent Systems*, 41–46.

Gatica-Perez, D. (in press). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*.

Hung, H., Babu, D., Ba, S., Odobez, J., & Gatica-Perez, D. (2008). Investigating automatic dominance estimation in groups from visual attention and speaking activity. Proceedings of the International Conference on Multimodal Interfaces ICMI), Chania, 233–236,

Hung, H., Jayagopi, D., Yeo, C., Friedland, G., Ba, S., Odobez, J., Ramchandran, K., Mirghafori, N., & Gatica-Perez, D. (2007). Using audio and video features to classify the most dominant person in a group meeting. Proceedings of the ACM International Conference on Mulmedia (ACM, MM), Augsburg, 835–838.

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., & Zhang, F. (2005). Automatic analysis of multimodal group actions in meetings. IEEE Transactions on Pattern Analysis and Machine Intelligence, *27*, 305–317.

Pentland, A. (2005). Socially aware computation and communication. *Computer, 38*, 33–40.

Rienks, R., & Heylen, D. (2005). *Automatic dominance detection in meetings using support vector machines*. Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh.

Fecha de recepción: 24 de septiembre de 2009
Fecha de aceptación: 29 de enero de 2010