Audiovisual Conflict Detection in Political Debates

Yannis Panagakis[1], Stefanos Zafeiriou[1], and Maja Pantic[1,2]

[1]Department of Computing,
Imperial College London,
180 Queens Gate,
London SW7 2AZ, U.K.

[2]EEMCS,
University of Twente,
Drienerlolaan 5,
7522 NB Enschede, The Netherlands

{i.panagakis,s.zafeiriou,m.pantic}@imperial.ac.uk

Abstract. In this paper, the automatic detection of conflict in audiovisual recordings of political debates is addressed. In contrast to the current state of the art in social signal processing, where only the audio modality is employed for analysing the human non-verbal behavior, we propose to use additionally visual features capturing certain facial behavioral cues such as head nodding, fidgeting and frowning which are related to conflicts. To this end, a dataset with video excerpts from televised political debates, where conflicts naturally arise, is introduced. The prediction of conflict level (i.e., conflict/nonconflict) is performed by applying the linear support vector machine and the collaborative representation-based classifier onto audio, visual, and audiovisual features. The experimental results demonstrate that the fusion of audio and visual features, outperform the accuracy in conflict detection, obtained by features that resort to a single modality (i.e., either audio or video).

1 Introduction

Social signals and social behaviors are the expression of one's attitude towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioral cues including facial expressions, body postures, gestures, and vocal outbursts like laughter. Social signals typically last for a short time (milliseconds, like turn taking, to minutes, like mirroring), compared to social behaviors that last longer (seconds, like agreement, to minutes, like politeness, to hours or days, like empathy) and are expressed as temporal patterns of non-verbal behavioral cues [1]. Since humans are predominantly social beings, the importance of social signals in everyday life situations is self-evident. In turn, multimedia data (e.g., television programs, movies, etc.) contain human social interactions and thus the automatic analysis and understanding of human social signals and social behaviors from audiovisual recordings is a cornerstone in the deployment of content-based multimedia indexing and retrieval, machine-mediated communication, state of the art human-computer interfaces, to mention but a few.

In spite of recent advances in social signal processing [1, 2] and machine analysis of relevant behavioral cues like blinks, smiles, head nods, laughter, and similar [3, 4, 2, 5], the research in machine analysis and understanding of more complex human social behaviors like interest, politeness, flirting, agreement, and conflict detection which this paper addresses, is still limited [6–8, 1, 2]. This can be partly attributed to both

1) an omnipresent neglect of the fact that observed behaviors may be influenced by those of an interlocutor and thus require analysis of both interactants at the same time, especially to measure such critically important patterns as mimicry, rapport, and disagreement, and 2) an overall lack of suitable annotated data that could be used to train the machine learning detectors for recognition of relevant phenomena [6, 1, 2]. Recent efforts in machine analysis of social interactions were aimed at analysis of various social signals including social dominance [9], engagement and hot-spots [10], behavioral codes (e.g., acceptance and blame) [11], and the analysis of personality [12]. These approaches employed statistical models trained on various lexical, prosodic and conversational features.

Conflict is used to label a range of human experiences, from disagreement to stress and anger, occurring when involved individuals act on incompatible goals, interests, or actions. Various research studies in human sciences argue that a “disagreement” does not have to result in a conflict; conflict describes a high level of disagreement, or “escalation of disagreement”, where at least one of the involved interlocutors feels emotionally offended. However, while conflict has been extensively investigated in human sciences and recognized as one of the main dimensions along which an interaction is perceived and assessed [13], machine analysis of conflicts is limited to automatic agreement/disagreement detection [6, 14–17] and is yet to be attempted based on audiovisual cues. To the best of our knowledge, the only work on the topic, and then based on audio cues only, is that by Kim *et al.* [7, 8], who investigated the degree of conflict in broadcasted political debates by employing various prosodic/conversational features.

This paper addresses the problem of conflict detection in videos. As opposed to Kim *et al.* [7, 8], the use of both audio and video modalities is investigated in conflict modeling and detection. Since, to the best of our knowledge, there are no available benchmarks datasets for audiovisual conflict detection, video excerpts from live political debates, where conflicts between participant naturally arise, are used. These videos have been extracted from more than 60 hours of live political debates, televised in between 2011 and 2012. In contrast with other benchmarks, political debates are real-world competitive multi-party conversations where participants do not act in a simulated context, but participate in an event that has a major impact on their real life (for example, in terms of results at the elections) [7]. Consequently, even if some constraints are imposed by the debate format, the participants have real motivations leading to real conflicts. From the entire dataset, 160 videos experts, with total duration 2h and 50 min, have been extracted. Theses videos have been annotated by 10 experts, in terms of continuous conflict intensity. The average annotation for each video is extracted by employing the Dynamic Probabilistic CCA [18]. Discrete labels (i.e., conflict/nonconflict here) are obtained next, by segmenting each video in non-overlapping conflict/nonconflict segments by applying an indicator function on the average annotation, resulting in 150 conflict (43 min) and 150 nonconflict (95 min) clips. The audio content of each video clip is parameterized in terms of prosodic and cepstral features, typically employed in affective computing [5]. Visually, the assessment of a conflict is highly related with the presence (or the absence) of certain facial behavioral cues such as head nodding, blinks, fidgeting and frowning [19]. To this end, the facial behavioral cues of each interactant are captured by tracking 68 facial points. The prediction of conflict level (i.e., conflict/nonconflict) is performed by applying a linear support vector machine (SVM) [20] and the collaborative representation-based classifier [21] onto feature vectors constructed by the audio modality, the video modality, and their combination. The experimental results indicate that the fusion of audio and video features outperforms the prediction accuracy obtained by features that resort to a single modality (i.e. either audio or video), yielding an accuracy of 85.59% when the collaborative representation-based classifier is employed in a two-class setting. Furthermore, the proposed method enables the modeling of conflict escalation and resolution.

The paper is organized as follows. In Section 2, the dataset and the annotation procedure is described. The audiovisual feature extraction process is outlined in Section 3. The experimental results are presented in Section 4. Conclusions are drawn in Section 5.

1.1 Dataset and Annotation Procedure

The dataset introduced in this paper for audiovisual conflict detection consists of video excerpts from televised political debates in Greek language. In particular, it consists of episodes of conflict escalation and resolution, which have been extracted from more than 60 hours of televised, live political debates aired as a part of the Anatropi Greek TV show[1]. Each debate includes at least two guests discussing under the moderation of the TV host.

From the entire dataset, 160 (170 min) non-overlapping dyadic episodes of conflict escalation have been manually extracted. For each episode of conflict, the database also contains an episode of conflict-free interaction of the two people in question. Each sample of the dataset is an audiovisual TV recording having both people involved in the dyadic episode in view. A sample frame from the dataset is depicted in Fig. 1. The episodes are of variable duration (i.e., 2 seconds to several minutes) and maybe noisy with a third party speaking in the background and people exhibiting large body movements.

The data have been annotated in terms of continuous conflict intensity by 10 expert annotators. The annotators assign a conflict intensity level, in the range $[0, 1]$, at each video frame by employing a joystick-based annotation tool, while they are watching each video excerpt in real time. They have been advised to annotate the videos by considering the physical (related to the behavior being observed) and inferential (related to the the interpretation of the discussion) layer of the conversation [7]. The physical layer includes the behavioral cues observed during conflicts and include interruptions, overlapping speech, cues related to turn-organization in conversations as well as but head nodding, fidgeting and frowning [19]. The inferential layer is based on the perception of the competitive processes [15] where conflict is considered as a “mode of interaction” where “the attainment of the goal by one party precludes its attainment by the others”. For instance, conflicting goals often lead to attempts of limiting, if not eliminating, the speaking opportunities of others in conversations.

To combine multiple annotators (Fig. 2(a)) subjective judgements, the Dynamic Probabilistic CCA with time warping [18] has been employed, yielding an average annotation for each video exert (Fig. 2(b)). The video excerpts are segmented next

[1] http://www.megatv.com/anatropi/

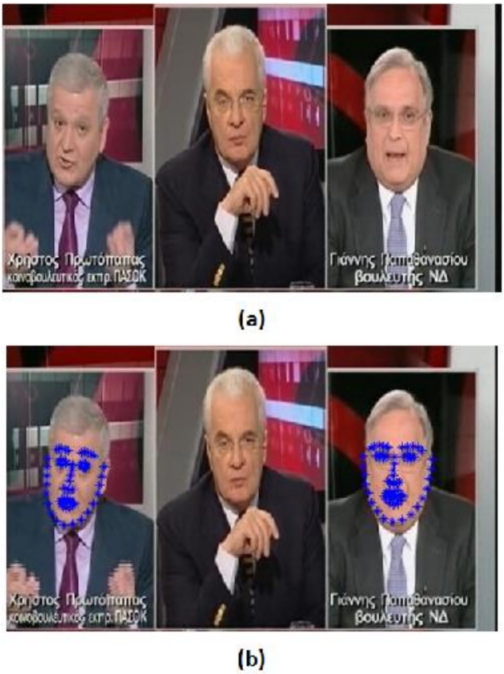
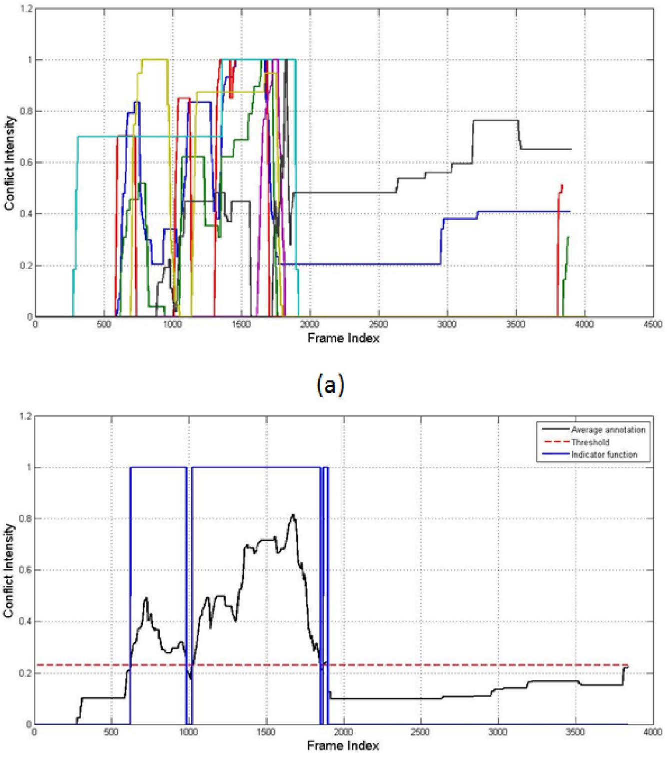


Fig. 1. (a) A sample snapshot from the dataset depicting the TV host and the two guests in conflict. (b) Facial points extracted from each guest, capturing the facial characteristics of the interactants being in conflict.

into non-overlapping conflict/nonconflict segments as follows: An indicator function assigns each frame the value 1 if the average annotation value is greater than its mean value and 0, otherwise. Segments corresponding to the discrete conflict/nonconflict sections of the video excerpt are depicted in Fig. 2(b). Finally, 150 conflict (43 min) and 150 nonconflict (95 min) clips, with discrete labels, have been selected. The annotated data will be available at `http://ibug.doc.ic.ac.uk/research`.



(a)

(b)

Fig. 2. (a) Conflict intensity as a continuous function of video frame index by various annotators. (b) Average continuous conflict intensity and segments corresponding to the discrete conflict/nonconflict sections of the video excerpt.

2 Feature Extraction

In this section the procedure followed for audiovisual feature extraction from each video excerpt in the dataset is outlined.

2.1 Audio Features

The audio content of each episode in the dataset is parameterized in terms prosodic and spectral features, namely the pitch related feature [22], the mean and the RSM energy feature, as well as the Mel-frequency cepstral coefficients (MFCCs) [23] and the Delta (differential) MFCCs.

The MFCCs [23] encode the frequency content of the speech signal by parameterizing the rough shape of spectral envelope and they have been successfully applied in Turn-taking analysis. Roughly speaking, the logarithm, which involved in the calculation of the MFCCs is a nonlinear transformation with additive property in the spectrum magnitude domain and thus the cepstral features can be consider as a superposition of latent variables, which are related to the speakers involved in the conversation. The MFCC calculation employs frames of duration 80 ms with a hop size of 40 ms, and a 42-band filter bank. The correlation between the frequency bands is reduced by applying the discrete cosine transform along the log-energies of the bands. The analysis yields a 23-dimensional vector of MFCCs for each video frame. This vector is appended with the Delta MFCCs, the 3 prosodic features, yielding an 49-dimensional audio feature vector for each video frame.

2.2 Visual Features

Cooper indicates that, facial behavioral cues related to conflict are head nodding, blinks, fidgeting, and frowning [19]. Consequently, the conflict can be visually captured by tracking the head pose, lips, eyebrows, eyelids, and related facial characteristics of the interactants in video sequences. To this end, the recently introduced persons' independent active appearance model, the so called active orientation model (AOM) [24] is employed for facial points tracking. In particular, the faces of the interactants are detected in the first frame of each video excerpt by the well-known Viola-Jones face detector [25]. Afterwards, the AOM is applied for tracking 68 2-dimensional facial points for each human throughout the video segment. As a result, for each video frame a 272-dimensional feature vector is obtained by stacking the points of each interactant. Facial points extracted from two interactants are depicted in Fig. 1 (b).

3 Experimental Results

In order to assess the performance of the proposed approach in conflict detection in political debates, experiments were conducted in the datset described in Section 2, by applying stratified 2-fold cross-validation.

To investigate the impact of each modality on conflict detection each video in the dataset is represented by three sequences of feature vectors. That is, by employing the 49-dimensional audio features, (audio modality), the 272-dimensional facial points (i.e., video modality) as well as the $272 + 49 = 321$-dimensional vector of audiovisual features. The latter feature vector is constructed by stacking the 49-dimensional audio on the top of the visual features for each video frame. Clearly, the length of the each feature sequence is equal to the number of the frames in video. The linear SVM [20] and the collaborative representation-based classifier (CRC) [21] are employed to assign each video frame into a class, namely to classify it as conflict or nonconflict. The classification results for frame level conflict detection are summarized in Table 1 for audio (A), video (V), and audiovisual features (AV). A single label for each video excerpt is obtained by averaging and rounding to the closest integer the predicted class labels of its frames. The classification results for video excerpt level conflict detection are summarized in Table 2.

Table 1. Frame-level conflict detection accuracy (%). The number within the parenthesis indicate the standard deviation.

Features	SVM	CRC
A	73.54 (0.31)	73.54 (0.31)
V	74.99 (0.31)	73.36 (0.31)
AV	78.58 (1.92)	79.95 (0.98)

Table 2. Video excerpt-level conflict detection accuracy (%). The number within the parenthesis indicate the standard deviation.

Features	SVM	CRC
A	73.76 (1.06)	74.59 (1.21)
V	82.92 (8.31)	83.92 (5.12)
AV	84.30 (10.60)	85.59 (2.91)

By inspecting Table 1 and Table 2, it is clear that the fusion of audio with visual features provide more accurate conflict prediction. In particular, the audiovisual feature discriminate the video exerts in those which contain conflicts and those which not contain conflicts with an accuracy of 85.59%, which is a significant improvement compared to that obtained by the audio features (i.e., 74.59%). This can be attributed to the fact that the audio channel is often noisy since a third party is speaking in the background. In contrast the video modality contain clear information about the behavior of the interactants.

Finally, there are indications that the conflict escalation and resolution can be modeled following the proposed approach, that is by classifying audiovisual features by the CRC. This can be done by assigning to each test video frame the average of the class labels within a window of 50 frames (i.e., 2 sec in our case). This maps the conflict intensity onto the continuous space. A demonstration of this can be found online[2], where the normalized in $[0, 1]$ conflict intensity level is depicted as a function of the video frame index. A snapshot of this demonstration is depicted in Fig. 3.

[2] http://youtu.be/yC9wrOA3RB0

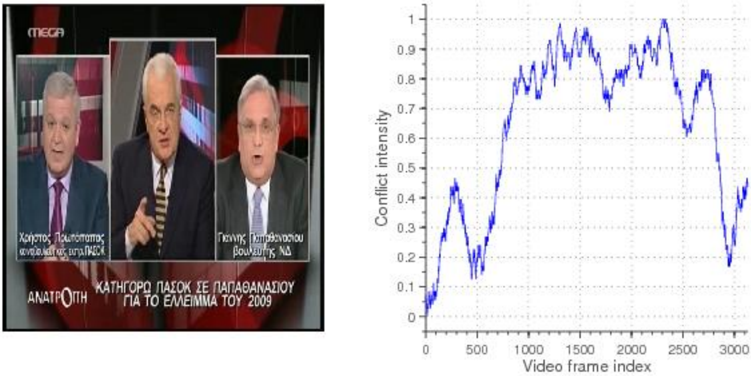


Fig. 3. (a) A sample snapshot from the dataset depicting the TV host and the two guests in conflict. (b) Conflict intensity as a function of video frame index.

4 Conclusions

In this paper, the problem of conflict detection in audiovisual recordings of political debated has been investigated. Audio and visual features have been demonstrated to detect the conflict more accurately than the features which resort to a single modality (i.e., either audio or video), when the CRC is employed.

In the future, the modeling of conflict escalation and resolution based on audiovisual and other features (e.g., conversational, lexical) will be investigated.

Acknowledgements

The work of Y. Panagakis was funded by the European Research Council under the FP7 Marie Curie Intra-European Fellowship. The work S. Zafeiriou was partially funded by the EPSRC project EP/J017787/1 (4DFAB). The work of M. Pantic was partially funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA).

References

1. Pantic, M., Cowie, R., D'ericco, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroder, M., Vinciarelli, A.: Social Signal Processing: The Research Agenda. Springer Verlag (2011)
2. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D"Errico, F., Schroeder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Trans. Affective Computing **3**(1) (2012) 69–87
3. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. J. Synthetic Emotion **1**(2) (2010) 68–99
4. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human-centred intelligent human-computer interaction ($hci2$): How far are we from attaining it? Int. J. Autonomous and Adaptive Communications Systems **1**(2) (2008) 168–187
5. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Analysis and Machine Intelligenc **31**(1) (2009) 39–58
6. Bousmalis, K., Morency, L., Pantic, M.: Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In: Proc. IEEE 2011 Int. Conf. Automatic Face and Gesture Recognition. (2011) 746–752
7. Kim, S., Valente, F., Vinciarelli, A.: Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In: Proc. 2012 IEEE Int. Conf. Audio, Speech and Signal Processing. (2012)
8. Kim, S., Yella, S.H., Valente, F.: Automatic detection of conflict escalation in spoken conversation. In: Proc. 13th Annual Conf. International Speech Communication Association. (2012)
9. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations from non-verbal activity cues,. IEEE Trans. Audio, Speech and Language Processing **17**(3) (2009) 501–513
10. Wrede, D., Shriberg, E.: Spotting hotspots in meetings: Human judgments and prosodic cues. In: Proc. Eurospeech. (2003) 2805–2808
11. Black, M., Katsamanis, A., Lee, C.C., Lammert, A., Baucom, B., Christensen, A., Georgiou, P., Narayanan, S.: Automatic classification of married couples' behavior using audio features. In: Proc. InterSpeech. (2010)
12. Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., Zancanaro, M.: Multimodal recognition of personality traits in social interactions. In: Proc. 2008 Int. Conf. Multimodal Interfaces. (2008) 253–60
13. Levine, J.M., Moreland, R.L.: Small groups. Oxford University Press (1998)
14. Bousmalis, K., Mehu, M., Pantic, M.: Towards the automatic detection of spontaneous agreement and disagreement based on non-verbal behaviour: A survey of related cues, databases, and tools. Image and Vision Computing Journal **31**(2) (2013) 203–221
15. M. Galley, K. McKeown, J.H., Shriberg, E.: Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. In: Proc. Meeting Association for Computational Linguistics. (2004) 669–676
16. Germesin, S., Wilson, T.: Agreement detection in multiparty conversation. In: Proc. Int. Conf. Multimodal Interfaces. (2009) 7–14
17. Hahn, S., Ladner, R., Ostendorf, M.: Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In: Proc. Human Language Technology Conf. of the NAACL. (2006) 53–56
18. Nicolaou, M.A., Pavlovic, V., Pantic, M.: Dynamic probabilistic cca for analysis of affective behaviour. In: Proc. 12th European Conference on Computer Vision, Florence, Italy (October 2012) 98–111

19. Cooper, V.W.: Participant and observer attribution of affect in interpersonal conflict: an examination of noncontent verbal behavior. J. Nonverbal Behavior **10**(2) (1986) 134144
20. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3) (2011) 1–27
21. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: Proc. 2011 Int. Conference on Computer Vision, Washington, DC, USA (2011) 471–478
22. Paul, B.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proc. of the Institute of Phonetic Sciences. (1993) 97110
23. Mueller, M., Ellis, D., Klapuri, A., Richard, G.: Signal processing for music analysis. IEEE J. Sel. Topics in Sig. Process. **5**(6) (2011) 1088–1110
24. Tzimiropoulos, G., Alabort, J., Zaferiou, S., Pantic, M.: Generic active appearance models revisited. In: Proc. 11th Asian Conf. Computer Vision. (2012)
25. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Computer Vision **57**(2) (2004) 137–154