The Conflict Escalation Resolution (CONFER) Database

Christos Georgakis, Yannis Panagakis, Stefanos Zafeiriou, Maja Pantic

PII: S0262-8856(16)30219-0
DOI: doi:10.1016/j.imavis.2016.12.001
Reference: IMAVIS 3584

To appear in: *Image and Vision Computing*

Received date: 15 June 2016
Revised date: 1 December 2016
Accepted date: 13 December 2016

The Conflict Escalation Resolution (CONFER) Database

Christos Georgakis[a,*], Yannis Panagakis[a], Stefanos Zafeiriou[a,b], Maja Pantic[a,c]

[a]*Department of Computing, Imperial College London, London, U.K.*
[b]*Center for Machine Vision and Signal Analysis, University of Oulu, Finland*
[c]*Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, The Netherlands*

Abstract

Conflict is usually defined as a high level of disagreement taking place when individuals act on incompatible goals, interests, or intentions. Research in human sciences has recognized conflict as one of the main dimensions along which an interaction is perceived and assessed. Hence, automatic estimation of conflict intensity in naturalistic conversations would be a valuable tool for the advancement of human-centered computing and the deployment of novel applications for social skills enhancement including conflict management and negotiation. However, machine analysis of conflict is still limited to just a few works, partially due to an overall lack of suitable annotated data, while it has been mostly approached as a conflict or (dis)agreement detection problem based on audio features only. In this work, we aim to overcome the aforementioned limitations by a) presenting the Conflict Escalation Resolution (CONFER) Database, a set of excerpts from audio-visual recordings of televised political debates where conflicts naturally arise, and b) reporting baseline experiments on audio-visual conflict intensity estimation. The database contains approximately 142 minutes of recordings in Greek language, split over 120 non-overlapping episodes of naturalistic conversations that involve two or three interactants. Subject- and session-independent experiments are conducted on continuous-time (frame-by-frame) estimation of real-valued conflict intensity, as opposed to binary conflict/non-conflict clas-

[*] Corresponding author.
E-mail address: christos.georgakis@imperial.ac.uk

sification. For the problem at hand, the efficiency of various audio and visual features and fusion of them as well as various regression frameworks is examined. Experimental results suggest that there is much room for improvement in the design and development of automated multi-modal approaches to continuous conflict analysis. The CONFER Database is publicly available for non-commercial use at `http://ibug.doc.ic.ac.uk/resources/confer/`.

Keywords: Automatic Conflict Analysis, Conflict Intensity Estimation, Conflict Detection, Behavioral Computing, Social Signal Processing, Behavioral Annotation

1. Introduction

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 intentions over resources or attitudes [1, 2]. 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. Similarly to other phenomena arising in social interactions [3, 4], conflict is largely manifested by means of non-verbal behavioral cues including facial expressions, body postures, gestures, and head movements, as well as conversational social signals including interruptions, overlapping speech, loudness and other cues associated with turn-organization [5]. Conflict, which has been recognized as one of the main dimensions along which a dyadic or multi-party social interaction is perceived, is usually accompanied by negative effects on communication and social life [6]. Hence, automatic analysis of conflict can be a cornerstone in the deployment of technologies targeting social interactions understanding and social skills enhancement such as content-based multimedia indexing and retrieval, machine-mediated communication, socially intelligent human-computer interfaces, to mention but a few.

Although conflict has been extensively investigated in human sciences, it has not received the same level of attention by the computing community. In spite of recent advances in social signal processing [7, 3, 4] and machine analysis of cues related to social behaviors [8, 9], research on machine analysis of conflict is still limited to just a few works that target automatic conflict detection based on audio features [10, 11, 12] or (dis)agreement detection [13, 14, 15]. This can be partially attributed to an overall lack of suitable annotated data that could be used to train the machine learning detectors for recognition of the relevant phenomena [14, 4]. Most importantly, given that interpersonal conflict is a mode of dyadic or multi-party interaction, automatic analysis of conflict is by itself a difficult task in terms of machine learning effort, since it requires the simultaneous analysis of more than one subjects at the same time. Also, the particularities of non-verbal communication due to conflictual conversation pose additional challenges to the related audio signal processing and computer vision tasks. For instance, interruptions and overlapping speech are more frequent when conflict takes place, which can largely affect the accuracy of speaker diarization or subsequent stages of audio feature extraction. When the visual modality is also considered, irregular postures or frequent and intense head and hand movements can lead to increased levels of visual noise pertaining to missing and incomplete data (e.g., partial image texture occlusions) or feature extraction errors (e.g., incorrect object localization, tracking errors).

Previous works on the automatic conflict analysis are characterized by the following main limitations.

- They are evaluated on corpora containing conversations that are captured in controlled, simulated conditions or on pre-segmented episodes of conflict/non-conflict.

- They are based exclusively on the audio modality (e.g., prosodic, conversational features), such as the works of Kim et al. [10, 11, 12], who investigated the degree of conflict in broadcasted political debates. The only audio-visual approach to conflict detection that we are aware of is [16], where robust, multi-modal fusion of audio-visual cues is utilized.

- They only deal with conflict detection or conflict escalation/resolution detection. These are approached as classification tasks aiming at estimating a single binary label (conflict/non-conflict) or discrete conflict intensity levels for the entire sequence or segments of it. The only work in the literature – that we are aware of – that has approached conflict in dimensional rather than categorical terms, i.e., as a continuous (real-valued) variable, and conflict intensity estimation as a regression task is [12].

In this paper, we provide a comprehensive description of the *Conflict Escalation Resolution (CONFER) Database*, a collection of audio-visual record-

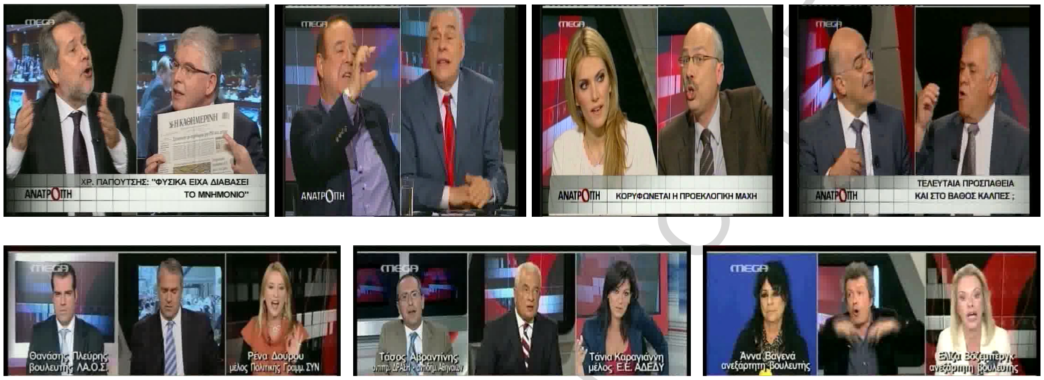
MEGA
Η ΚΑΘΗΜΕΡΙΝΗ
ΧΡ. ΠΑΠΟΥΤΣΗΣ: "ΦΥΣΙΚΑ ΕΙΧΑ ΔΙΑΒΑΣΕΙ ΤΟ ΜΝΗΜΟΝΙΟ"
ΑΝΑΤΡΟΠΗ
ΚΟΡΥΦΩΝΕΤΑΙ Η ΠΡΟΕΚΛΟΓΙΚΗ ΜΑΧΗ
ΤΕΛΕΥΤΑΙΑ ΠΡΟΣΠΑΘΕΙΑ ΚΑΙ ΣΤΟ ΒΑΘΟΣ ΚΑΛΠΕΣ ;
Θανάσης Πλεύρης βουλευτής ΛΑ.Ο.Σ.
Ρένα Δούρου μέλος Πολιτικής Γραμμ. ΣΥΝ
Τάσος Αβραντίνης
Τάνια Καραγιάννη μέλος Ε.Ε. ΑΔΕΔΥ
Άννα Βαγενά ανεξάρτητη βουλευτής
ανεξάρτητη βουλευτής

Figure 1: Characteristic frames from episodes of the Set *two* (top row) and *three* (bottom row) of the CONFER Database.

ings of naturalistic interactions from political debates suitable for the investigation of conflict behavior. These recordings have been manually extracted from more than 60 hours of live political debates, televised in Greece between 2011 and 2012. In contrast with other corpora, 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) [10]. Consequently, even if some constraints are imposed by the debate format, the participants have real motivations leading to real conflicts.

From the entire dataset, 120 video excerpts have been extracted from a total of 27 TV broadcasts, with total duration amounting to approximately 142 minutes. The dataset is split into 2 sets, namely *two* and *three*, which consist of recordings containing interactions that involve two or three participants, respectively. All 120 videos have been annotated by 10 experts, in terms of continuous conflict intensity. The CONFER Database has been partially presented at previous works (see [17, 16, 18]), but a complete description of the data and available annotations, has not been reported so far. The database is publicly available for non-commercial use at `http://ibug.doc.ic.ac.uk/resources/confer/`. Along with the audio-visual episodes and the annotations, audio and visual features (facial tracking points and SIFT) are also provided (see Section 4).

This work is novel not only in providing a comprehensive description of this database, which is suitable for the investigation of conflict behavior in naturalistic conversations, but also in reporting baseline experiments that could serve as a benchmark for efforts in the field. These experiments primarily aim to overcome the last two of the aforementioned limitations of previous works on automatic conflict analysis, namely by i) examining both audio and visual features as well as fusion of them for the target problem, and ii) addressing *continuous-time* (frame-by-frame) estimation of *continuous-valued* conflict intensity. For each Set of the database, we conduct two baseline experiments in which the efficiency for the problem at hand of various visual (shape- and appearance-based) descriptors and audio features as well as fusion of them, and classifiers, respectively, is examined. A cross-validation experimental scenario is employed in order to assess performance of the baseline predictive frameworks on collectively all audio-visual recordings of the CONFER Database. A challenging experimental protocol is established with all experiments being subject- and session-independent. This is to ensure that the sequences used for testing involve different subjects from different TV broadcasts compared to those used in the training phase.

Overall, the contributions of this paper are as follows.

- A comprehensive description of the Conflict Escalation Resolution (CONFER) Database is provided. The audio-visual data and available annotations are described in detail.

- The presented baseline experiments constitute the first *audio-visual* approach in the literature to *continuous-time* (frame-by-frame) estimation of *continuous-valued* conflict intensity.

- This database contains naturalistic, competitive conversations from political debates where conflict naturally arises, and, as such, is primarily intended for research targeting automatic conflict analysis. However, it could also be a valuable source for studies of other social phenomena such as (dis)agreement, interest, mimicry, turn-taking, and back channel (i.e., head nods) communication.

- The provided audio-visual episodes of conflict have been filmed in real-world "in-the-wild" conditions involving a wide range of views, amenable lighting conditions, spontaneous and overlapping speech, and abrupt head and body movements or occlusions. Hence, it can be also exploited to evaluate the robustness of signal processing and machine learning techniques for automatic speech recognition and speaker identification, face detection and facial point tracking, head pose estimation, to mention but a few.

The remainder of the paper is as follows. Section 2 provides a list of existing databases that have already been or could be used for automatic analysis of conflict and similar social signals and phenomena. Section 3 presents in detail the audio-visual data and conflict intensity annotations included in the CONFER Database. Section 4 describes the methodology employed for the baseline experiments on continuous conflict intensity estimation that are presented in Section 5, while Section 6 concludes the paper.

2. Prior Work

The only existing database that has been released primarily to serve research on machine analysis of conflict is the SSPNet Conflict Corpus [20], which consists of 1430 clips of 30 seconds extracted from the Canal9 Corpus [19] – a collection of audio-visual recordings from 45 political debates aired on the Swiss TV (in French) – corresponding to 138 subjects in total. Each clip of the database has been annotated in terms of a single continuous conflict score in the range $[-10, +10]$ for the purposes of the sequence-level binary classification and regression tasks of the Conflict Sub-Challenge included in the Interspeech 2013 Computational Paralinguistics Challenge [20]. Pesarin et al. [28] have manually segmented 13 debates from the SSPNet Conflict Corpus, with a total duration of 6 h and 27 min, into conflictual and non-conflictual intervals for conflict detection. Recently, Kim et al. [12] have relied on Mechanical Turk crowdsourcing to have the corpus annotated in terms of continuous (real-valued) conflict intensity, using two separate questionnaires, one for the *physical* layer and the other for the *inferential* layer of the conversation (see Section 3 for definitions). However, the annotations of both works mentioned above constitute a sequence-level rather than a frame-by-frame characterization of conflict.

Audio-visual recordings of political debates have been recently utilized for research on detection of the similar behavioral phenomenon of

[1] The Green Persuasive Database can be found online at `http://sspnet.eu/2009/12/green-persuasive-database/`.

[2] The SEWA Database is available online at `http://db.sewaproject.eu/`.

Table 1: Summary of the databases that have or could be used for automatic analysis of conflict as well as other social behaviors and signals.

Database	**# Subjects**	**Duration**	**Audio-visual?**	**Synchronous?**	**Cultural Background**
Canal 9 [19]	190	43 h 10 min	✔	✔	Swiss (French-speaking)
SSPNet Conflict Corpus [20] (subset of Canal 9)	138	11 h 55 min	✔	✔	Swiss (French-speaking)
[13] (subset of Canal 9)	28	not known	✔	✔	Swiss (French-speaking)
AMI [21]	213	100 h	✔	✔	Mostly non-native English speakers
AMIDA [22]	not known	10 h	✔	✔	Mostly non-native English speakers
ICSI [23]	53	75 h	✗	–	28 native English speakers (mostly American), the rest non-native (12 German)
Green Persuasive[1]	16	not known	✔	✔	not known
Wolf [24]	36	7h	✔	✔	Mostly non-native English speakers
Mission Survival [25]	44	6h	✔	✔	Canadian
MAHNOB Mimicry [26]	60	11h	✔	✔	Spanish, French, Greek, English, Dutch, Portuguese, Romanian
SEWA[2]	398	34h 35 min	✔	✔	British, German, Hungarian, Greek, Serbian, Chinese
[27]	208	62 h 48 min	✔	✔	77% Caucasian, 8% Afr. American, 5% Asian or Pac. Islander, 5% Latino(a), 1% Native American, 4% Other

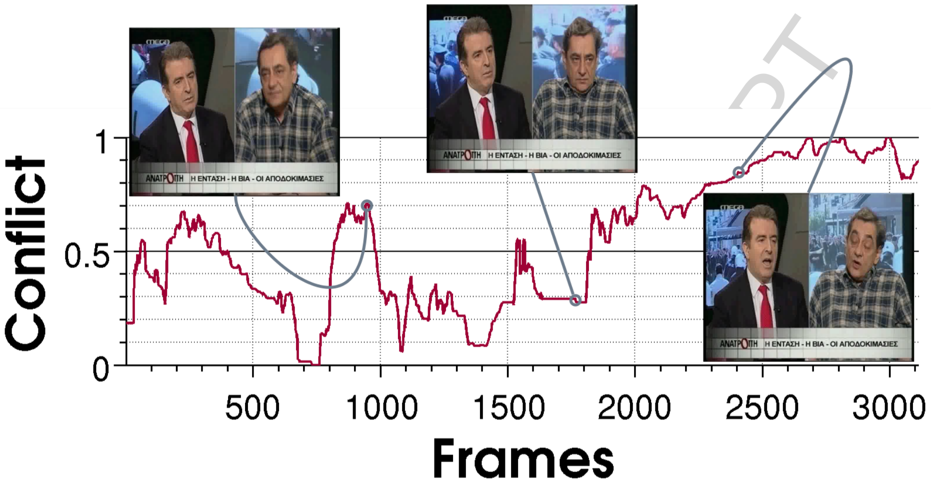

Figure 2: Conflict intensity annotations along with three characteristic frames shown for the sequence *20120326_ seq3* from the Set *two* of the CONFER Database.

(dis)agreement [13] (see [29]) for a survey). The latter has been also investigated by means of experiments performed on meeting corpora such as the AMI Corpus [21] and the ICSI Corpus [23]. Other databases, albeit not annotated in terms of (dis)agreement or conflict, that contain multiple instances of the latter behaviors as well as other social behaviors (e.g., interest, politeness, mimicry, flirting), social signals (e.g., social dominance, engagement, hot-spots, acceptance, blame) and personality traits include [22, 24, 25, 27], the Green Persuasive Dataset and the newly released SEWA Database. It is worth mentioning that the SEWA Database is the largest and the richest database of human conversational and emotional behaviour that has been released so far. Table 1 provides a concise summary of the existing databases that have already or could be used for automatic analysis of conflict and similar social signals and phenomena.

3. Database

In this section, we provide a comprehensive description of the CONFER Database, a collection of audio-visual recordings of naturalistic interactions from political debates.

Data. The database consists of video excerpts from televised political debates in Greek language. In particular, it contains episodes of conflict escalation and resolution, which have been extracted from more than 60 hours of live political debates aired as a part of the Anatropi Greek TV show[3]. Each debate includes at least two guests discussing under the moderation of the TV host.

From the entire collection of the TV programme broadcasts, 120 non-overlapping episodes of conflict escalation have been manually extracted. These audio-visual excerpts are divided into two Sets, which are balanced in terms of total duration, namely the Set *two* that contains 73 episodes of dyadic interactions, and the Set *three* that contains 47 episodes of interactions among three subjects. Overall, these episodes correspond to a total duration of approximately 142 minutes and to a total number of 54 subjects, 43 male and 11 female. It is worth mentioning that the episodes contain debates that may have more than one instances of conflict escalation, yet they always end with conflict resolution. For all recordings, the video stream has been recorded at 25 frames per second, while the sample rate of the audio channel is 22050 Hz. Each video sequence of the dataset has a spatial resolution of 720×576 pixels and has all participants involved in the episode in view. The duration of the episodes varies from 20.2 seconds to 534.0 seconds, having a mean and standard deviation of 71.0 and 70.5 seconds, respectively, as computed for the whole dataset. Characteristic frames from the dataset are depicted in Fig. 1.

Due to the spontaneous and competitive nature of the interactions contained in the CONFER Database, various types and levels of noise are incurred in the data. Regarding the audio channel, speaker diarization and speech recognition are rendered difficult since the interlocutors often interrupt or talk over one another, driven by anger or agitation or aiming to dominate the dialogue. In some of the recordings, a third party speaking in the background is involved. Also, in most of the cases speech is emotionally colored and thus often fragmented and disorganized or extremely rapid and even unintelligible.

Regarding the visual stream, camera angles can vary a lot across episodes or even within the same episode, while illumination conditions vary less. Depending on the way the interlocutors are positioned in the studio, the former are often portrayed at large head pose *pan* angles or even in almost-profile view, due to them looking at their interlocutor rather than the camera fixed on them. Moreover, due to the involved parties being engaged in naturalistic

[3] `http://www.megatv.com/anatropi/`.

competitive conversations, the subjects often perform abrupt and extreme head movements (e.g., head nods, shakes, tilts), body movements (e.g., forward/backward leaning, spinning periodically on their swivel chairs) and gestures (e.g., hand crosses, hand wags). The aforementioned conditions pose obstacles to the computer vision pre-processing tasks, such as face detection, facial point tracking and registration [30, 31], since the latter have to cope with frequent and large out-of-plane head rotations and occlusions [32, 33].

Annotations. The data have been annotated on a frame-by-frame basis in terms of continuous (real-valued) conflict intensity by 10 expert annotators, all of them being native Greek speakers. The annotation task is carried out in real time, i.e., while the annotators are watching each audio-visual excerpt, by employing a joystick-based annotation tool. The tool records the conflict intensity level in the continuous range $[0, 1000]$ at a variable sampling rate, which is approximately 64 samples per second in average. All annotations are subsequently down-sampled to the video frame rate of 25 frames per second. The procedure followed so as to extract a single ground truth annotation sequence from the 10 available annotations for each episode of the CONFER Database is described in detail in Section 4.2. Ground truth annotations of conflict intensity are plotted as a function of time for a sequence of the database along with three characteristic frames in Fig. 2.

The annotators have been advised to annotate the videos by considering the *physical* (related to the behavior being observed) and the *inferential* (related to the interpretation of the discussion) layer of the conversation [10]. 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 head nodding, fidgeting and frowning [5]. The *inferential layer* is based on the perception of the competitive processes occurring in conversations where conflict is viewed as a 'mode of interaction' governed by the principle that *"the attainment of the goal by one party precludes its attainment by the others"* [2, 15]. For instance, conflicting goals often lead to attempts of limiting, if not eliminating, the speaking opportunities of others in conversations. In view of the demanding nature of the task of annotating conflict in real time and in terms of both conversational layers, all annotators were initially 'trained' on a small subset (∼10%) of the CONFER Database episodes. In particular, they were instructed to watch these episodes as many times as they considered necessary and retain the annotation that best assessed conflict intensity in terms of both layers. For each of the remaining episodes of the database, the annotators were allowed two plays, and again the most suitable annotation was retained.

4. Methodology

In this section, the methodology employed for the baseline experiments conducted on the CONFER Database for audio-visual continuous-time conflict intensity estimation is described.

4.1. Sets and Protocol

Two baseline experiments are conducted for each Set of the CONFER Database. A *subject- and session-independent cross-validation* experimental protocol is employed. Specifically, each Set is divided in 5 segments, balanced in terms of duration, containing videos that include different interactants and have been broadcast at different times. In each fold, 3 segments are used for training, one for validation (parameter tuning) and the remaining one for testing, and the average value over all test sequences of each evaluation metric (see Section 4.5) is retained. The process is repeated 5 times, until all episodes have been used for testing. Finally, the mean and standard deviation of the metrics, as computed over all 5 folds, are reported.

4.2. Annotations

Recent studies on combining multiple annotations of human behavior or affect have provided evidence suggesting that the average of multiple annotations can lie far away from the actual ground truth and thus lead to ill-generalizable models [34]. This is mainly due to the subjectivity of annotators and the variability related to their age and gender or their stress, fatigue, attention or even intention while annotating (e.g., there can be *spammer* annotators that they do not even pay attention during the annotation process). Furthermore, when the task in question is temporal, additional noise in the set of multiple annotations is entailed by the temporal lags in the perception and annotation of the related events.

Motivated by the aforementioned findings, herein we follow a supervised approach to fusing the multiple available annotations. Specifically, Canonical Correlation Analysis (CCA) [35] is employed for each sequence to extract subspaces that are maximally correlated for the set of 10 annotations available and the corresponding audio-visual feature set. For all experiments presented in this paper, the coefficient corresponding to the first component

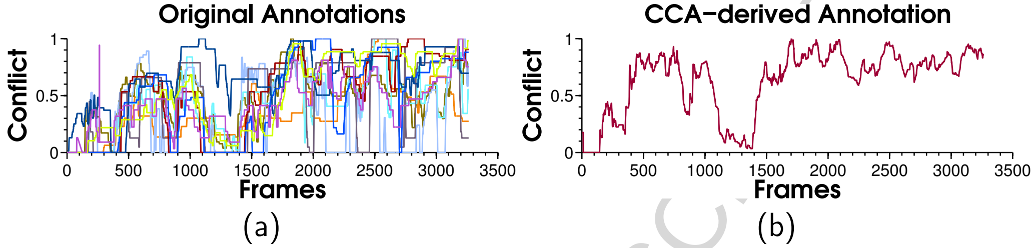

Figure 3: Annotations illustrated as a function of time for the sequence *20120206_ seq5* from the Set *three* of the CONFER Database. (a) Original annotations from 10 annotators rescaled in $[0, 1]$, and (b) Ground truth annotations derived by performing CCA on the original annotations and the corresponding features.

of the CCA-derived annotation subspace is used as the ground truth annotation for each episode. The latter is rescaled in the continuous range $[0, 1]$. Original annotations of conflict intensity from the 10 annotators as well as the CCA-derived annotation for a sequence of the CONFER Database are plotted as a function of time in Fig. 3.

4.3. Features

The various audio and visual features as well as fusion of them that are used in the experiments of this study are described in what follows.

Audio features. As mentioned above, most of the existing approaches to automatic conflict analysis have relied almost exclusively on audio features [10, 11, 12] such as spectral, prosodic, durational, lexical and turn organization descriptors. A concise review of audio-based approaches to (dis)agreement and conflict detection is provided in [12].

In this work, we employ the openSMILE feature extractor [36] to obtain the ComParE acoustic feature set of 65 low-level descriptors (LLD) (4 energy-related, 55 spectral and 6 voicing-related), which has been successfully applied for automatic recognition of paralinguistic phenomena [37]. The 65 LLD used are summarized in Table 3 in [38]. The audio features extracted for each sequence of the CONFER Database are down-sampled to 25 Hz frequency to match the frame rate of the video stream. Similarly to [39, 37] the audio features of each sequence are z-normalized (each feature component is normalized to mean=0 and standard deviation=1).

Visual features. In recent years, research in behavioral and affective computing as well as signal processing has gradually shifted from audio-only (or even video-only) systems to audio-visual approaches [8, 9]. As a matter of fact, the latter have been shown to outperform uni-modal frameworks in various related tasks such as continuous interest prediction [40, 16], detection of behavioral mimicry [41], and dimensional and continuous affect prediction [39], to mention but a few. Notably, other challenging problems such as accent classification [42, 43, 44] and pain intensity estimation [45] have been addressed based exclusively on visual features.

Motivated by the aforementioned works and deviating from a common practice in automatic conflict analysis where only audio features are employed (e.g., [10, 11, 12]), in this paper we utilize also visual features for conflict intensity estimation. Our aim is to capture facial behavioral cues that are deemed intrinsically correlated with conflict, such as smiling, blinking, head nodding, flouncing and frowning [5, 14]. Both shape- and appearance-based descriptors are examined. Note that the video stream of each episode from the CONFER Database is spatially cropped at each frame so that a separate video stream is obtained for each one of the participants involved in the conversation. The Menpo project [46] has been employed in this study for all visual feature extraction tasks, which are described as follows.

Facial point tracking: First, 68 fiducial facial points are detected at each frame of each cropped video sequence portraying a single interactant. To this end, we employ the Gauss-Newton Deformable Part Model in [47], which when combined with a person-specific face detector produces very accurate results [48]. The coordinates of 49 facial landmarks are retained for each frame by excluding the facial points that correspond to the face boundaries. Next, the effects of head translation, scale and in-plane rotation are removed by universally aligning the tracking points with the 'mean' shape computed over all frames through a 2-D non-reflective similarity transformation.

Shape features: Principal Component Analysis (PCA) [49] is applied on the aligned tracking points to yield a low-dimensional shape descriptor for each frame. In particular, the coordinates of the 49 facial landmarks are projected onto the subspace spanned by the 'eigenshapes' of a pre-trained Active Shape Model (ASM) [50]. The latter has been previously trained on collectively 4 datasets of faces "in-the-wild", and thus its principal components efficiently 'explain' variations of shape corresponding, for instance, to out-of-plane rotations, different face anatomy characteristics and subtle expression-related deformations. For each video frame, 18 coefficients that account for 95% of the total variance are retained for each subject. The final feature vector for each frame is obtained by concatenating the descriptors for

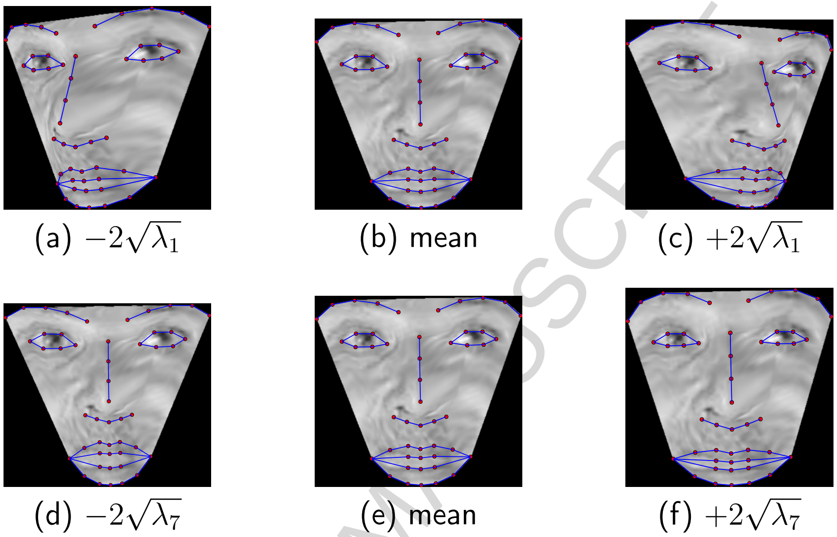

Figure 4: Effect on the mean shape ((b), (e)) of varying the 1st ($i = 1$) component (pose-related) and the 7th ($i = 7$) component (expression-related) of the Active Shape Model used herein for shape feature extraction by $-2\sqrt{\lambda_i}$ and $2\sqrt{\lambda_i}$, where λ_i denotes the respective eigenvalue.

all interactants in the episode.

Inspired by [51], we follow a face-anatomy-driven rather than a simply data-driven approach to identifying the most suitable feature representation of facial shape for the problem at hand. To this end, we visually inspect the deformation pattern associated with each component of the ASM. We observe that the first 6 components capture head movements (rigid face motion), while the remaining 12 components capture expression-related deformations (non-rigid face motion). The discriminative power of both pose- and expression-related shape features as well as the combination of them – which we henceforth call *Pose*, *Expression* and *Points*, respectively – is investigated for the target problem. In Fig. 4, one can see the mean shape and the effect on it of varying the 1st ($i = 1$) component (pose-related) and the 7th ($i = 7$) component (expression-related) by $-2\sqrt{\lambda_i}$ and $+2\sqrt{\lambda_i}$, where λ_i denotes the variance explained by the respective component. It is evident that the former component is associated with out-of-plane head rotation (*yaw*), whereas the latter component is associated with deformations related to sadness/happiness (*frown*/*smile*).

Appearance features: Previous frameworks targeting biometrics and affective computing tasks such as face recognition [52] and pain intensity estimation [45] have relied on appearance features locally extracted from a pre-defined grid of rectangular regions in face images registered in frontal pose. However, this technique is not suitable for databases including images that portray faces with large head pose angles, as is the case with the CONFER Database, since the 2D registration process unavoidably induces pixel artefacts and texture discontinuities. Furthermore, some researchers are critical of the grid-based feature extraction, suggesting that the sub-regions are not necessarily well aligned with meaningful facial features [53].

Motivated by these findings and other recent works [54, 33, 55], in this study we adopt a hybrid approach to appearance feature extraction. In particular, we first apply the same transformation used for point registration to the pixel intensities of each face image to remove translation, scale and in-plane rotation effects. Subsequently, features are extracted from the intensities lying within rectangular regions (patches) of dimension 20×20 pixels centered at each facial point. Facial point tracking and point/image registration results are depicted for each interlocutor in Fig. 5 for 2 characteristic frames from a sequence of the CONFER Database.

Two appearance-based descriptors are examined herein, namely *Scale-Invariant Feature Transform (SIFT)* [56] and *Discrete Cosine Transform (DCT)* [57]. SIFT is a rotation- and scale-invariant descriptor that captures local orientation information in images, while DCT is a frequency-based descriptor that projects pixel intensities onto real cosine basis functions. For SIFT, we extract a 4×4 array of 8-bin orientation histograms for each image patch. For DCT, the two-dimensional DCT is employed and the first 128 out of the zig-zag-arranged coefficients, which correspond to the lowest frequencies, are retained, so that the final dimensionality matches that of SIFT. For both descriptors, the features calculated from the total of 49 patches are concatenated into a single vector. For each frame, the final representation is formed by concatenating the feature vectors for all interlocutors (two or three). Finally, dimensionality is reduced in a supervised manner, by applying CCA on the features and corresponding annotations. The CCA coefficients of the feature set corresponding to 95% of the total energy are retained, thus resulting in dimensionality of 75 (65) and 63 (56) for the Set *two* (*three*) for SIFT and DCT, respectively. Note that all visual features are ℓ_2-normalized.

Fusion. To investigate which features carry complementary information with regards to manifestations of conflict in conversational and emotional

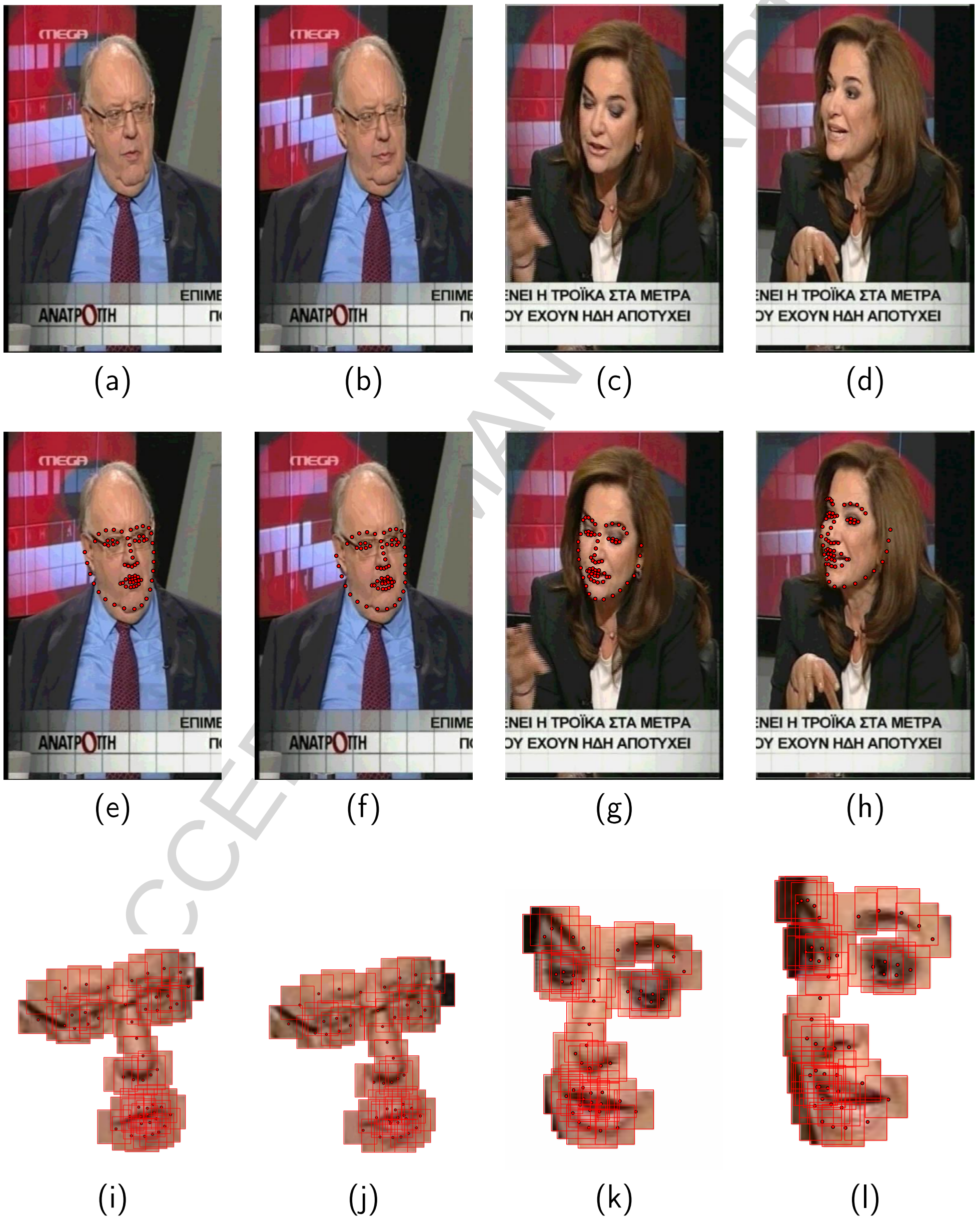

Figure 5: Tracking and point/image registration results shown for each subject for 2 characteristic frames (frame 683 and frame 762) of the sequence *20111212_ seq1* from the Set *two* of the CONFER Database. (a)-(d) Original video frames, (e)-(h) Original tracking points superimposed on the original video frames, and (i)-(l) Rectangular patches extracted around the aligned points on the aligned video frames.

behavior data and thus could help improve performance of conflict analysis tools, in this study we examine also *feature-level fusion*. Both *intra-modality* (Video-Video) and *inter-modality* (Audio-Visual) fusion is investigated. For the former case, we combine the expression-related shape descriptor with each appearance descriptor, that is, *Expression+SIFT* and *Expression+DCT* as well as the *Points* descriptor (the whole shape-based feature vector) with the appearance descriptor that performs best in the first baseline experiment (see Section 5.1). This is done on the feature level, that is, by concatenating the respective feature vectors.

For audio-visual (AV) fusion, we follow a more sophisticated approach, motivated by recent evidence which suggests that feature-level AV fusion can be sub-optimal and highly problematic mainly due to (i) the two modalities being recorded at different measurement and temporal scales and (ii) the detrimental effect of increased dimensionality on the classifier's performance [8]. To overcome the aforementioned limitations, we perform CCA to derive linear, maximally correlated components among the audio and visual feature sets. After retaining the components that account for the 95% of energy for each of the sets, the resulting CCA coefficients are concatenated to form the final AV feature representation. Note that for AV fusion, audio features are combined only with the best-performing out of the (single- or multi-feature) visual descriptors examined in the first baseline experiment (see Section 5.1).

4.4. Classifiers

Four classifiers that have been extensively used for temporal modeling of human behavior and affect are examined, namely *Support Vector Regression (SVR)* [58], *Random Forests for Regression (RF)* [59], *Continuous Conditional Random Fields (CCRF)* [60], and *Long-Short Term Memory (LSTM) Neural Networks* [61]. LIBSVM [62], `scikit-learn` [63], [60], and the CUda RecurREnt Neural Network Toolkit (CURRENNT) [64] are used to train SVR, RF, CCRF and LSTMs, respectively. For each fold of the cross-validation experiments, the validation set is used to optimize the classifiers in terms of Correlation (COR) for SVR, RF and CCRF (see Section 4.5), and Root Mean Squared Error (RMSE) for LSTMs[4].

[4] CURRENNT [64] only supports RMSE criterion for the objective function of LSTMs.

SVR [65] is a discriminative regression framework that extends Support Vector Classification (SVC) to the continuous (real-valued) targets, and is one of most commonly used regressors in the fields of affective computing and social signal processing [9, 3] with applications to various tasks such as continuous and dimensional emotion prediction [39], and social signal/behavior (e.g., laughter/conflict) detection/recognition [20], to mention but a few. In this study, linear SVR with ϵ-insensitive loss function is examined, whose parameters are optimized by means of a suitable grid search. In particular, the regularization parameter C is optimized in the set $\{10^{-5}, 10^{-4}, \ldots, 1\}$, the convergence tolerance parameter *tol* in the set $\{10^{-5}, 10^{-4}, \ldots, 10^{-2}\}$, while for the ϵ parameter 50 values logarithmically spaced in the range $[10^{-2}, 1]$ are examined.

RF [59] is an ensemble learning algorithm that combines unpruned Decision Tree learners based on random split selection of feature subspaces. RF have gained popularity in recent years within the computer vision and machine learning communities (e.g., [66, 67]) as they combine the ability to handle large training datasets with computational efficiency and good generalizability. The two most critical parameters in the RF design, that is the number of trees T in the forest and the number of features F selected to split each node, are optimized in the range $T \in \{100, 500, 1000, 2000\}$ and $F \in \{\sqrt{p}, p/3, p/2\}$, respectively, where p denotes the dimensionality of the feature vector.

CCRF [60] is an undirected graphical model-based discriminative framework that extends the traditional Conditional Random Fields (CRF) [68] to the case of continuous (real-valued) output. CCRF have been applied in combination with SVR for the task of continuous and dimensional emotion prediction [60]. Herein, we follow the approach in [60] in using linear SVR (exactly as described above) to learn the *vertex* (static) features of the graphical model and ten *edge* (temporal) features, that is, 5 neighbor $n = \{1, 2, \ldots, 5\}$ and 5 distance similarities $\sigma = \{2^6, 2^7, \ldots, 2^{11}\}$ (see [60] for details).

LSTMs [69] constitute an extension of the traditional Recurrent Neural Network architecture that is efficient in capturing contextual statistical regularities with large and unknown lags in time-series data. LSTMs have been successfully applied to various behavioral and affective computing tasks such as continuous and dimensional affect prediction [70, 39], visual-only accent classification [43], and audio-visual depression scale prediction [71]. Herein, we use bi-directional LSTMs with 1 hidden layer of 128 memory blocks. The output layer consists of a single node whose sigmoid-function activation is used as the estimate of the conflict intensity. The networks are trained with stochastic gradient descent with a batch size of 5 sequences for a maximum of 1000 epochs. Finally, zero-mean Gaussian noise of variance 0.1 is added to the features and early stopping is employed to prevent overfitting.

4.5. Evaluation Metrics

Performance is measured for each test sequence based on two metrics, namely the *Pearson's Correlation coefficient (COR)* and the *Intra-class Correlation Coefficient (ICC)* [72]. Both metrics are computed for each test sequence, and the average value over all test sequences is retained for each fold. Finally, the mean and standard deviation of each metric over all 5 folds are reported.

The Pearson's Correlation coefficient (COR) is, along with the Mean Squared Error (MSE), the most commonly used evaluation metric in the affective computing literature [9, 39]. We have opted to use COR in this study over MSE since the former can capture linear structural information about how ground truth annotations and predictions vary together through the calculation of the covariance [9]; if the two measurements have a perfect linear relationship, then COR becomes 1 (complete positive relationship) or -1 (complete negative relationship). This property of the correlation is deemed advantageous for the experimental setting of our study that deals with continuous-time (frame-by-frame) estimation of conflict intensity.

The Intra-class Correlation Coefficient (ICC) [72], initially proposed as a metric for rater reliability in behavioral measurements, has been recently applied in providing a measure of 'consistency' or 'agreement' between ground truth annotations of behavioral or affective attributes provided by humans and corresponding predictions yielded by automated approaches (e.g., [45, 73]). It typically expresses the fraction of the total variance across all ratings and subjects (including random error in the 'judgements') 'explained' by the component of variance due to the targets alone [72]. Herein, we employ the coefficient ICC(3,1), which corresponds to the scenario *'Each target is assessed by each rater, with a single measurement being available for each rater and the raters being the only raters of interest'* [72]. For each automated framework examined, the ICC is computed based on the ground truth annotations and the predicted values of conflict intensity.

To obtain a 'human' baseline ICC result, i.e., a measure of 'level of consistency amongst 10 humans in assessing conflict intensity', we also compute the ICC amongst the 10 available annotations for each sequence. This facilitates a more fair evaluation of the various automated approaches examined in the experiments presented in Section 5.2. In particular, it enables us to compare the degree of conformity – in ICC terms – between the 'mean annotation' and the conflict intensity predictions yielded by the various frameworks to the degree of conformity amongst the measurements of conflict intensity obtained by 10 humans for the same data. The mean (standard deviation) of the 'inter-annotator' ICC is 0.495 (0.037) for the Set *two* and 0.414 (0.057) for the Set *three*, respectively.

5. Results

In this section, experimental results are reported and discussed separately for each of the two baseline experiments conducted on the CONFER Database for audio-visual continuous-time conflict intensity estimation.

5.1. Baseline Experiment I: Feature Comparison

In the first experiment of this study, we investigate the efficiency of the various audio and visual (shape- and appearance-based) descriptors as well as the (Video-Video and Audio-Visual) fusion of them described above, for the task of *continuous-time (frame-by-frame) estimation of continuous (real-valued) conflict intensity*. In total, 10 features (incl. fusion) are examined, namely *Audio*, *Pose*, *Expression (Expr.)*, *Points*, *SIFT*, *DCT*, *Expr.+SIFT* , *Expr.+DCT*, *Points+SIFT*, and *Fusion (AV)*. For the regression stage of this experiment, we use linear SVR which is one of the most commonly used regression frameworks in the literature for dimensional behavior and affect modeling [9]. We first examined the single-feature systems. Then, for Video-Video fusion, we chose to examine the combination of the whole shape feature vector (*Points*) with the best-performing appearance descriptor, i.e., *SIFT*, hence *Points+DCT* is not considered. Finally, for audio-visual fusion, we examined the combination of *Audio* features with the best-performing out of all visual features and fusion of them, i.e., *Expr.+SIFT*.

Conflict intensity estimation results, in terms of COR averaged over all 5 folds of the cross-validation experiment, are shown in the bar graph of Fig. 6 for the Sets *two* and *three* of the CONFER Database. Among the single-feature frameworks, the best performance of COR = 0.233 and COR = 0.302 for the Set *two* and *three* is achieved by *Audio* and *SIFT*, respectively. Note that *Audio* is the only feature that accounts for lower performance on the Set *three* than on the Set *two*, presumably due to the increased number of speakers in the former case incurring a larger number of speaker diarization errors. On the other hand, for all visual features there is a large discrepancy between the performances achieved on the Sets *two* and *three*, with the latter being higher in all cases. This finding makes sense upon observing that it is often the case with the recordings of the Set *three* that not all interactants are recorded in the same studio and thus some of them retain a (quasi-)frontal view during the conversation by looking straight at the camera rather than at their interlocutors. Under these conditions, the computer vision tasks of facial point tracking and image registration are rendered much easier and hence accurate, thus leading to more efficient and error-free visual feature extraction.

Among shape features, *Pose* features largely outperform *Expression* features, with the latter leading to a rather poor performance when considered alone. This conforms to recent evidence [14, 29] suggesting that head gestures (e.g., head nod, shake, roll, 'cut-off') are among the most common non-verbal cues through which (dis)agreement is manifested, hence the efficiency of head pose features in capturing the latter and conflict as well. Also, the poor performance yielded by *Expression* can be partially attributed to the high variation of expression-related facial deformations in the CONFER Database, which entails that a lot of the latter do not convey conflict information and thus are uninformative for the task at hand.

Appearance features perform more accurately than shape features. This is exactly as expected; while shape features are capable of capturing coarse deformations related to facial expression, appearance features are efficient in encapsulating finer movements and tale-telling transient features such as bulges, wrinkles and furrows [8, 74, 44]. Also, SIFT outperforms DCT. This is again not a surprising result given that SIFT features extracted from local patches around facial landmarks have been shown to be efficient for automatic face analysis "in-the-wild" [54]. Also, DCT being less efficient than SIFT in this experiment can be partially attributed to its Fourier-based transformation, which is applied locally, capturing energy characteristics in the visual scene which are unrelated to conflict (e.g., uninformative facial expressions, illumination changes caused by head movements). It is also worth mentioning that, the shape-based *Expression* descriptor, despite performing poorly when used in isolation, leads to performance improvement when combined with either of the appearance descriptors. This behavior can be explained considering that *Expression* captures coarse non-rigid deformations from the whole face which are complementary to the local subtle movements encoded

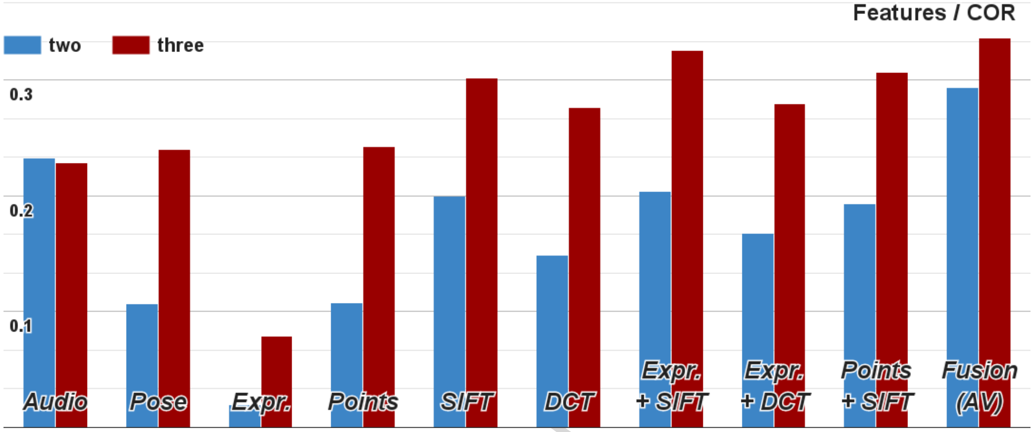

Figure 6: Baseline Experiment I: Conflict intensity estimation results, in terms of COR averaged over all 5 folds, as obtained by linear SVR trained with the various visual and audio features as well as fusion of them (Video-Video and Audio-Visual) examined herein, for the Sets *two* and *three* of the CONFER Database.

by the appearance descriptors extracted from the local patches.

Finally, audio-visual fusion outperforms all remaining frameworks, leading to COR = 0.294 and COR = 0.336 for the Set *two* and *three*, respectively. This result provides a strong indication that behavioral patterns associated with continuous-in-time manifestations of conflict under unconstrained recording conditions are more accurately recognized when cues from both the audio and video modality are considered, as is the case with other social behaviors such as (dis)agreement, mimicry, interest, and flirting [3, 4].

5.2. Baseline Experiment II: Classifier Comparison

In the second experiment of this study, we investigate the efficiency of the various classifiers described above in modeling and predicting conflict intensity in continuous time for each test sequence, approached again as a regression problem on a frame-by-frame basis. The features utilized to train the classifiers are those that performed best in the previous experiment, i.e., *Audio* and *Expr.+SIFT* for Audio and Video, respectively, and *Audio+Expr.+SIFT* for audio-visual (AV) fusion.

Conflict intensity estimation results, in terms of the COR and ICC metrics averaged over all 5 folds of the cross-validation experiment, are reported in Table 2a and Table 2b for the Set *two* and *three* of the CONFER Database,

Table 2: Baseline Experiment II: Conflict intensity estimation results, in terms of COR and ICC averaged over all 5 folds, as obtained by each feature-classifier combination examined herein for the Sets (a) *two*, and (b) *three* of the CONFER Database, respectively. The corresponding standard deviation values are reported inside parentheses. The best COR and ICC performances for the uni- and multi-modal frameworks (A: Audio, V: Visual, AV: Audio-Visual) are shown in boldface.

Classifier / Feature	**Audio (A)**		**Expr.+SIFT (V)**		**Fusion (AV)**	
	COR	**ICC**	**COR**	**ICC**	**COR**	**ICC**
SVR	0.233 (0.064)	**0.774** (0.031)	0.204 (0.090)	0.174 (0.030)	**0.294** (0.065)	**0.781** (0.029)
RF	0.170 (0.054)	0.144 (0.020)	0.052 (0.024)	0.168 (0.042)	0.178 (0.053)	0.160 (0.020)
CCRF	**0.285** (0.177)	0.160 (0.355)	0.026 (0.067)	-0.001 (0.000)	0.221 (0.075)	0.163 (0.357)
LSTMs	0.232 (0.092)	0.178 (0.033)	0.126 (0.071)	0.183 (0.070)	0.251 (0.070)	0.195 (0.065)

(a) Set *two*

Classifier / Feature	**Audio (A)**		**Expr.+SIFT (V)**		**Fusion (AV)**	
	COR	**ICC**	**COR**	**ICC**	**COR**	**ICC**
SVR	0.229 (0.063)	**0.687** (0.036)	**0.326** (0.076)	0.357 (0.144)	**0.336** (0.033)	**0.296** (0.187)
RF	0.156 (0.061)	0.213 (0.050)	0.158 (0.108)	0.204 (0.092)	0.173 (0.092)	0.198 (0.089)
CCRF	0.213 (0.036)	0.045 (0.044)	0.153 (0.130)	0.014 (0.031)	0.211 (0.109)	0.014 (0.023)
LSTMs	0.259 (0.068)	0.221 (0.077)	0.185 (0.100)	0.186 (0.045)	0.148 (0.055)	0.195 (0.022)

(b) Set *three*

respectively. The best performances of COR = 0.294 and COR = 0.336 for the Set *two* and *three*, respectively, are those achieved by audio-visual fusion in the previous experiment. Interestingly, both the aforementioned best-performing frameworks employ SVR in the regression stage. However, it is worth noting that not for all classifiers does fusion result in improved performance (in terms of COR) over that furnished by the corresponding uni-modal systems. This can be partially attributed to different classifiers being to a different degree sensitive to (i) gross errors and outliers in the audio or/and the video stream which, in turn, result in erroneous estimates of the correlated components obtained by the classical CCA due to its reliance on least squares minimization, and (ii) errors induced by feature pro- and post-processing (e.g., normalization, AV synchronization). A partial remedy to the above-mentioned limitations could be sought in either applying more robust techniques for the extraction of individual and correlated components, such as the one proposed in [16], or ‘delegating’ both the tasks of modeling each stream separately and uncovering the correlations between them to the classifier by means of *model-level fusion* (see [8] for a survey of different types of fusion).

Regarding the uni-modal frameworks, the best performances of COR = 0.285 and COR = 0.326 are accounted for by the combination of *Audio* with CCRF and *Expr.+SIFT* with SVR for the Set *two* and *three*, respectively. The superiority of SVR among classifiers for the multi-modal frameworks and the high accuracy achieved by it also when trained with features from a single modality conforms to previous evidence indicating its robustness to overfitting and suitability for continuous prediction of behavior and affect dimensions [20, 39]. CCRF also yield accurate predictions in this experiment, presumably thanks to their ability to learn the conflict ‘history’ across successive observations of continuous conversational data given that they, like their discrete-output counterpart (CRF), relax the assumption of conditional independence of the features [68]. We argue that their performance for conflict intensity prediction could be improved by (i) examining different functions for the *vertex* and *edge* features (e.g., non-linear regressor for the *vertex* features), and (ii) investigating different normalization schemes, to which they have shown to be quite sensitive [70]. LSTMs trained with *Audio* features also achieve high COR values for both Sets, albeit on par with or not much higher than those achieved by SVR. This result might seem counter-intuitive at first sight, since LSTMs, similarly to CCRF, are capable of capturing long-range dependencies between successive observations and, as such, have been shown successful in continuous modeling of human behavior and affect [70, 39, 71]. However, we argue that the relatively low performance of LSTMs in this experiment is mainly due to them having been trained based on RMSE and that, by using an alternative implementation that allows COR-based objective function for LSTMs training, one will most probably achieve much higher performance. The same holds for the RF frameworks which have been also trained on the basis of mean-squared generalization error and thus are agnostic to contextual temporal information. The poor performance of RF for this experiment can be also attributed to the random feature selection process employed to determine the split at each node; this practice can result in sub-optimal partitioning of the feature space, especially in the case of insufficient training data [75]. To alleviate this limitation, one could resort to a semi-supervised approach to node splitting, such as the one proposed in [75], that is, to use also unlabeled data to guide the node splitting.

As for the results in terms of ICC, SVR combined with *AV Fusion* and

Audio accounts for the best performances of ICC = 0.781 and ICC = 0.687 for the Set *two* and *three*, respectively. It is worth noting that the best ICC scores obtained by *Audio* are much higher than those obtained by the visual descriptor *Expr.+SIFT*. In other words, the predictions yielded by the former framework are much more 'consistent' in terms of ICC with the ground truth annotations than those yielded by the latter. This behavior can be partially attributed to the annotation process. In particular, it is highly likely that the annotators, who are all native speakers of Greek that is the language spoken in the CONFER Database, relied much more on the audio modality while annotating since in that alone they could easily identify informative cues associated with conflict escalation/resolution in terms of both the *physical* layer (e.g., interruptions, overlapping speech) and the *inferential layer* (e.g., sarcasm, rudeness, confrontation) of the conversation. The impact of this condition is larger in absolute terms for the ICC rather than the COR metric in the results reported in Table 2. This is presumably due to the random error associated with the 'raters' decreasing significantly for the audio-based system and thus leading to an increase in the ratio of variances to which ICC equals (see Section 4.5 and [72] for more details).

Furthermore, it is also worth noting that the aforementioned best performances in terms of ICC exceed the corresponding values of ICC = 0.495 and ICC = 0.414 measured amongst the 10 annotators for the Set *two* and *three*, respectively. This signifies that the corresponding frameworks, which were trained using the 'mean annotator' annotations, learned the trend of the 'mean annotator' better and were able to reproduce the trend accurately. This result is quite encouraging in that it reveals that even uni-modal systems based on a commonly used classifier can be more 'consistent' with the 'mean human rating' in assessing conflict intensity than several humans are with one another on the same dataset.

Overall, the relatively low results achieved in both experiments described above can be attributed to (i) the challenging nature of the CONFER Database, which consists of spontaneous conversational data where conflict naturally arises, (ii) the demanding subject- and session-independent experimental protocol adopted in this study, and (iii) the abundance of the data (106536 and 106404 frames in total for the Set *two* and *three*, resp.), which are all tested by means of cross-validation. However, these results indicate that there is much room for improvement for tools targeting the task at hand. We hope that these findings will encourage further research in the future in the development of audio-visual approaches to automatic analysis of conflict as well as similar behavioral and affective phenomena.

6. Conclusion

In this paper, we presented the Conflict Escalation Resolution (CONFER) Database, a set of audio-visual recordings of naturalistic interactions from political debates suitable for the investigation of conflict behavior. The database contains 142 minutes of recordings in total and is the first of its kind to have been annotated in terms of continuous (real-valued) conflict intensity on a frame-by-frame basis. Data and annotations are publicly available for non-commercial use at `http://ibug.doc.ic.ac.uk/resources/confer/`. The CONFER Database can be used for the development of tools that target automatic analysis of conflict or other social behaviors (e.g., (dis)agreement, interest, politeness) and social signals (e.g., social dominance, engagement, hot-spots), automatic speech recognition, recognition of non-verbal behavioral cues (e.g., facial expressions, body postures, gestures, and vocal outbursts) as well as related audio processing and computer vision tasks (e.g., speaker diarizers, facial point trackers). In this study, we also reported benchmark results of subject- and session-independent experiments by means of which the efficiency of commonly used audio and visual features and fusion of them as well as classifiers was examined for continuous-time estimation of continuous conflict intensity. These experiments represent the first systematic study of automatic conflict analysis viewed as a frame-by-frame regression task, with results indicating that there is much room for improvement.

Acknowledgments

This work was funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). Yannis Panagakis was also partially funded by the ERC under the FP7 Marie Curie Intra-European Fellowship. The work of Stefanos Zafeiriou was funded by the FiDiPro program of Tekes (project number: 1849/31/2015).

References

[1] J. Allwood, Cooperation, competition, conflict and communication, Gothenburg Papers in Theoretical Linguistics 94 (2007) 1–14.

[2] C. M. Judd, Cognitive effects of attitude conflict resolution, Journal of Conflict Resolution 22 (3) (1978) 483–498.

[3] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, M. Schroeder, Bridging the gap between social animal and unsocial machine: A survey of social signal processing, IEEE Transactions on Affective Computing 3 (1) (2012) 69–87.

[4] M. Pantic, A. Vinciarelli, Social Signal Processing, Springer, 84–93, 2014.

[5] V. W. Cooper, Participant and observer attribution of affect in interpersonal conflict: an examination of noncontent verbal behavior, Journal of Nonverbal Behavior 10 (2) (1986) 134–144.

[6] J. M. Levine, R. L. Moreland, Small groups: key readings, Psychology Press, 2008.

[7] M. Pantic, R. Cowie, F. D'ericco, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroder, A. Vinciarelli, Social Signal Processing: The Research Agenda, 511–538, 2011.

[8] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (1) (2009) 39–58.

[9] H. Gunes, B. Schuller, Categorical and dimensional affect analysis in continuous input: Current trends and future directions, Image and Vision Computing 31 (2) (2013) 120–136.

[10] S. Kim, F. Valente, A. Vinciarelli, Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),, 5089–5092, 2012.

[11] S. Kim, S. H. Yella, F. Valente, Automatic detection of conflict escalation in spoken conversations., in: INTERSPEECH, 1167–1170, 2012.

[12] S. Kim, F. Valente, M. Filippone, A. Vinciarelli, Predicting Continuous Conflict Perception with Bayesian Gaussian Processes, IEEE Transactions on Affective Computing 5 (2) (2014) 187–200.

[13] K. Bousmalis, L. P. Morency, M. Pantic, Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition, in:

IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG), 746–752, 2011.

[14] K. Bousmalis, M. Mehu, M. Pantic, Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools, in: IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops, 1–9, 2009.

[15] M. Galley, K. McKeown, J. Hirschberg, E. Shriberg, Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 669, 2004.

[16] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, M. Pantic, Robust Correlated and Individual Component Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Special Issue in Multimodal Pose Estimation and Behaviour Analysis, (accepted) .

[17] Y. Panagakis, S. Zafeiriou, M. Pantic, Audiovisual Conflict Detection in Political Debates, in: Computer Vision-ECCV 2014 Workshops, Springer, 306–314, 2014.

[18] L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, S. Nikitidis, M. Pantic, Probabilistic Slow Features for Behavior Analysis, IEEE transactions on neural networks and learning systems 27 (5) (2016) 1034–1048.

[19] A. Vinciarelli, A. Dielmann, S. Favre, H. Salamin, Canal9: A database of political debates for analysis of social interactions, in: International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII), 1–4, 2009.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al., The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism .

[21] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al., The AMI meeting corpus, in: International Conference on Methods and Techniques in Behavioral Research, vol. 88, 2005.

[22] J. Carletta, Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus, Language Resources and Evaluation 41 (2) (2007) 181–190.

[23] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, et al., The ICSI meeting corpus, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, I–364, 2003.

[24] H. Hung, G. Chittaranjan, The idiap wolf corpus: exploring group behaviour in a competitive role-playing game, in: ACM International Conference on Multimedia, 879–882, 2010.

[25] F. Pianesi, M. Zancanaro, B. Lepri, A. Cappelletti, A multimodal annotated corpus of consensus decision making meetings, Language Resources and Evaluation 41 (3-4) (2007) 409–429.

[26] S. Bilakhia, S. Petridis, A. Nijholt, M. Pantic, The MAHNOB Mimicry Database: A database of naturalistic human interactions, Pattern recognition letters 66 (2015) 52–61.

[27] P. G. Georgiou, M. P. Black, A. C. Lammert, B. R. Baucom, S. S. Narayanan, “That’s Aggravating, Very Aggravating”: Is It Possible to Classify Behaviors in Couple Interactions Using Automatically Derived Lexical Features?, in: Affective Computing and Intelligent Interaction, Springer, 87–96, 2011.

[28] A. Pesarin, M. Cristani, V. Murino, A. Vinciarelli, Conversation analysis at work: detection of conflict in competitive discussions through semi-automatic turn-organization analysis, Cognitive processing 13 (2) (2012) 533–540.

[29] K. Bousmalis, M. Mehu, M. Pantic, Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools, Image and Vision Computing 31 (2) (2013) 203–221.

[30] C. Sagonas, Y. Panagakis, S. Zafeiriou, M. Pantic, Robust Statistical Face Frontalization, in: Proceedings of IEEE Int’l Conf. on Computer Vision (ICCV 2015), Santiago, Chile, 2015.

[31] C. Sagonas, Y. Panagakis, S. Zafeiriou, M. Pantic, Robust Statistical Frontalization of Human and Animal Faces, International Journal of Computer Vision, Special Issue on "Machine Vision Applications" .

[32] G. Papamakarios, Y. Panagakis, S. Zafeiriou, Generalised Scalable Robust Principal Component Analysis, in: British Machine Vision Conference (BMVC 2014), 2014.

[33] C. Georgakis, Y. Panagakis, M. Pantic, Discriminant Incoherent Component Analysis, IEEE Transactions on Image Processing 25 (5) (2016) 2021–2034.

[34] M. A. Nicolaou, V. Pavlovic, M. Pantic, Dynamic Probabilistic CCA for Analysis of Affective Behaviour and Fusion of Continuous Annotations, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7) (2014) 1299–1311.

[35] T. W. Anderson, An introduction to multivariate statistical analysis, Tech. Rep., Wiley New York, 1962.

[36] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the ACM international conference on Multimedia (ACMMM), 1459–1462, 2010.

[37] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, K. Scherer, On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common., Frontiers in psychology 4 (2012) 292–292.

[38] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, B. Schuller, Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data, Pattern Recognition Letters 66 (2015) 22–30.

[39] M. A. Nicolaou, H. Gunes, M. Pantic, Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space, IEEE Transactions on Affective Computing (2011) 92–105.

[40] M. A. Nicolaou, Y. Panagakis, S. Zafeiriou, M. Pantic, Robust Canonical Correlation Analysis: Audio-visual fusion for learning continuous interest, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1522–1526, 2014.

[41] S. Bilakhia, S. Petridis, M. Pantic, Audiovisual Detection of Behavioural Mimicry, in: Affective Computing and Intelligent Interaction (ACII 2013), Geneva, Switzerland, 123–128, 2013.

[42] C. Georgakis, S. Petridis, M. Pantic, Visual-only discrimination between native and non-native speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4828–4832, 2014.

[43] C. Georgakis, S. Petridis, M. Pantic, Discriminating Native from Non-Native Speech Using Fusion of Visual Cues, in: ACM International Conference on on Multimedia (ACMMM), Orlando, Florida, USA, 1177–1180, 2014.

[44] C. Georgakis, S. Petridis, M. Pantic, Discrimination Between Native and Non-Native Speech Using Visual Features Only, IEEE Transactions on Cybernetics (TCYB) (99) (2015) 1–14.

[45] S. Kaltwang, S. Todorovic, M. Pantic, Doubly Sparse Relevance Vector Machine for Continuous Facial Behavior Estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (to appear).

[46] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, S. Zafeiriou, Menpo: A Comprehensive Platform for Parametric Image Alignment and Visual Deformable Models, in: Proceedings of the ACM International Conference on Multimedia (ACMMM), Orlando, Florida, USA, 679–682, 2014.

[47] G. Tzimiropoulos, M. Pantic, Optimization Problems for Fast AAM Fitting in-the-Wild, in: IEEE International Conference on Computer Vision (ICCV), 2013.

[48] G. Chrysos, E. Antonakos, S. Zafeiriou, P. Snape, Offline Deformable Face Tracking in Arbitrary Videos, in: Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15), Santiago, Chile, 2015.

[49] I. Jolliffe, Principal component analysis, Wiley Online Library, 2002.

[50] T. Cootes, E. Baldock, J. Graham, An introduction to active shape models, Image processing and analysis (2000) 223–248.

[51] S. Petridis, M. Pantic, Audiovisual discrimination between speech and laughter: Why and when visual information might help, IEEE Transactions on Multimedia 13 (2) (2011) 216–234.

[52] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in: Computer vision-ECCV 2004, Springer, 469–481, 2004.

[53] D. Huang, C. Shan, M. Ardabilian, Y. Wang, L. Chen, Local binary patterns and its application to facial image analysis: a survey, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 41 (6) (2011) 765–781.

[54] E. Antonakos, J. Alabort-i medina, S. Zafeiriou, Active Pictorial Structures, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 5435–5444, 2015.

[55] C. Sagonas, Y. Panagakis, S. Zafeiriou, M. Pantic, RAPS: Robust and Efficient Automatic Construction of Person-Specific Deformable Models, in: IEEE Conference on Computer Vision & Pattern Recognition (CVPR), 2014.

[56] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[57] G. Potamianos, A. Verma, C. Neti, G. Iyengar, S. Basu, A cascade image transform for speaker independent automatic speechreading, in: IEEE International Conference on Multimedia and Expo (ICME), 1097–1100, 2000.

[58] A. J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and computing 14 (3) (2004) 199–222.

[59] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[60] T. Baltrusaitis, N. Banda, P. Robinson, Dimensional affect recognition using continuous conditional random fields, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1–8, 2013.

[61] A. Graves, Rnnlib: A recurrent neural network library for sequence learning problems, 2013.

[62] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3) (2011) 27.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[64] F. Weninger, J. Bergmann, B. Schuller, Introducing currennt: The munich open-source cuda recurrent neural network toolkit, The Journal of Machine Learning Research 16 (1) (2015) 547–551.

[65] V. Vapnik, S. E. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, in: Advances in neural information processing systems (NIPS), Citeseer, 1996.

[66] B. Schuller, Z. Zhang, F. Weninger, G. Rigoll, Using multiple databases for training in emotion recognition: To unite or to vote?, in: INTERSPEECH, Citeseer, 1553–1556, 2011.

[67] H. Chen, J. Li, F. Zhang, Y. Li, H. Wang, 3D model-based continuous emotion recognition, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1836–1845, 2015.

[68] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data .

[69] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18 (5) (2005) 602–610.

[70] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies., in: INTERSPEECH, vol. 2008, Citeseer, 597–600, 2008.

[71] L. Chao, J. Tao, M. Yang, Y. Li, Multi task sequence learning for depression scale prediction from video, in: IEEE International Conference on Affective Computing and Intelligent Interaction (ACII), 526–531, 2015.

[72] P. E. Shrout, J. L. Fleiss, Intraclass correlations: uses in assessing rater reliability., Psychological bulletin 86 (2) (1979) 420.

[73] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, AVEC 2016-Depression, Mood, and Emotion Recognition Workshop and Challenge, arXiv preprint arXiv:1605.01600 .

[74] M. Pantic, Facial Expression Recognition, 1–8, 2014.

[75] X. Liu, M. Song, D. Tao, Z. Liu, L. Zhang, C. Chen, J. Bu, Semi-supervised node splitting for random forest construction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 492–499, 2013.

Highlights

- The Conflict Escalation Resolution (CONFER) Database is presented.
- CONFER contains 142 minutes (120 episodes) of recordings in Greek language.
- Episodes are extracted from TV political debates where conflicts naturally arise.
- Experiments are the first approach to continuous estimation of conflict intensity.
- Performance of various audio and visual features and classifiers is evaluated.