

Non-Verbal Communication Analysis in Victim-Offender Mediations

Víctor Ponce-López, Sergio Escalera, Marc Pérez, Oriol Janés, and Xavier Baró

Abstract—In this paper we present a non-invasive ambient intelligence framework for the semi-automatic analysis of non-verbal communication applied to the restorative justice field. In particular, we propose the use of computer vision and social signal processing technologies in real scenarios of Victim-Offender Mediations, applying feature extraction techniques to multi-modal audio-RGB-depth data. We compute a set of behavioral indicators that define communicative cues from the fields of psychology and observational methodology. We test our methodology on data captured in real world Victim-Offender Mediation sessions in Catalonia in collaboration with the regional government. We define the ground truth based on expert opinions when annotating the observed social responses. Using different state-of-the-art binary classification approaches, our system achieves recognition accuracies of 86% when predicting satisfaction, and 79% when predicting both agreement and receptivity. Applying a regression strategy, we obtain a mean deviation for the predictions between 0.5 and 0.7 in the range [1-5] for the computed social signals.

Keywords—*Victim-Offender Mediation, Multi-modal Human Behavior Analysis, Face and Gesture Recognition, Social Signal Processing, Computer Vision, Statistical Learning.*

I. INTRODUCTION

RESTORATIVE justice is an international social movement for the reform of criminal justice. This approach to justice focuses on the needs of the victims, who take an active role in the process, while offenders are encouraged to take responsibility for their actions *to repair the harm they have done* [1]. It differs, therefore, with respect to the classic justice framework, where justice seeks to satisfy abstract legal principles or to directly punish offenders. One of the common procedures offered to victims is the possibility of exchanging their impressions with a mediator, in a program known as the Victim-Offender Mediation (henceforth VOM) program. Given the sensitive nature of the cases, the process consists initially of a set of individual encounters, where each party involved (i.e., victim or offender) attends an interview or meeting with a mediator to analyze the problem in depth. The decision is

V. Ponce-López is with the IN3 at the Open University of Catalonia, the Dept. MAiA at the University of Barcelona, and the Computer Vision Center at the Autonomous University of Barcelona.

S. Escalera is with the Dept. MAiA at the University of Barcelona, and the Computer Vision Center at the Autonomous University of Barcelona.

M. Pérez is a B.S. student in Computer Science and Mathematics at the Faculty of Mathematics. He has collaborated in this work with the Dept. MAiA at the University of Barcelona.

O. Janés obtained his B.S. in Computer Science at the Faculty of Mathematics. He has collaborated in this work with the Dept. MAiA at the University of Barcelona.

X. Baró is with the EIMT at the Open University of Catalonia, and the Computer Vision Center at the Autonomous University of Barcelona.

then taken as to whether the victim and the offender might engage in a joint encounter. Figure 1 (a) shows an example of a real VOM scenario.

In the VOM process, the goal is to reach a restitution agreement by seeking to balance the interests of each of the parties, conditioned of course by the events that have occurred and the legal proceedings. This agreement can be reached in one of two ways. First, there are pre-conditioning factors to a case, given its particular facts, which make mediation feasible or not. Second, high levels of agreement and expressed satisfaction between the parties and the mediator are indicators of whether the VOM process is likely to end in success or failure [2]. The emergence of these indicators depends on a large set of factors that are not only concerned with the professionalism of the mediator, but are also related to other factors including the applicability of mediation, the participants' traits, human relationships, and first impressions, among others. Furthermore, if we examine each of the participants (victim, offender, and mediator), certain characteristics, including their cultural background, education, and social status, are likely to have a high impact on the success or otherwise of the process [3], [4]. For example, feelings of inferiority may emerge during the conversations if the mediator employs higher language skills than those employed by one or other of the parties involved. In such cases, the mediation process is perhaps more susceptible to failure than it is to success. Thus, it is important to emphasize the particular behavioral traits of those involved in a VOM from a psychological perspective.

Participant roles are clearly defined in these conversational processes, as they are in similar scenarios, such as job interviews. The mediator explains the process and listens to the other parties, maintaining his or her impartiality at all times, whereas the victim and offender are more concerned with protecting their own interests and may appear quite wrapped up in the problem they face. In this sense, no standard guidelines exist for establishing the best course of actions or identifying the psychological mechanisms for achieving the desired mediation goals. Moreover, the mediators have often received a different type of training focused neither in psychology nor in observational methodology, two fields concerned with understanding how biological, psycho-social, and environmental factors can impact participants' moods. However, if these techniques were studied and established as self-knowledge tools for both communication and intervention, mediators could obtain valuable feedback from the mediations and guide the process towards success.

In this context, multi-modal intelligent systems can be used to achieve these objectives. This information can be analyzed for different modules in order to extract features from each

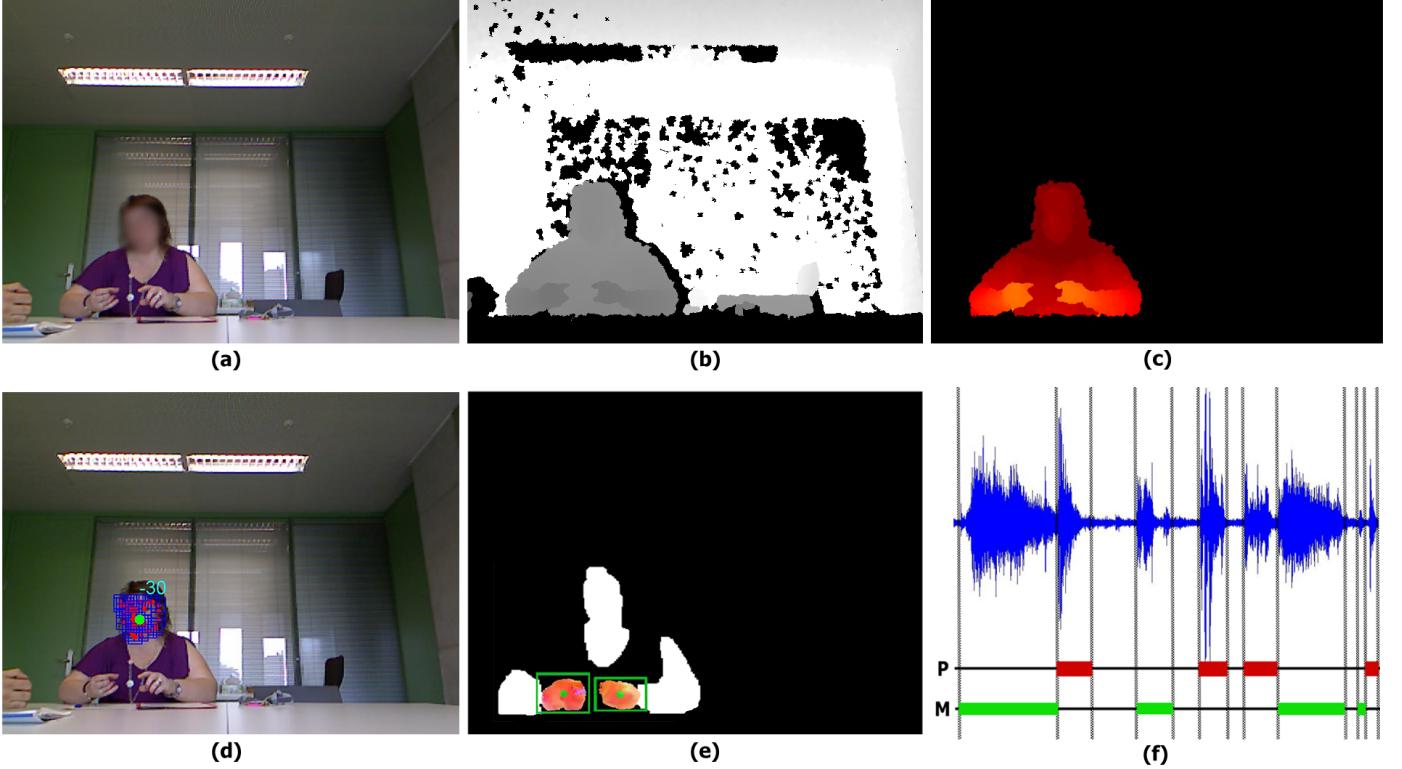


Fig. 1. Examples of the multi-modal feature extraction². Images (a) and (b) are the RGB and depth images, respectively. Image (c) shows the upper body obtained from the Random Forest user segmentation. In image (d), both face detection and head pose estimation are shown. Hand segmentation is shown in image (e). Across the regions segmented by color, the optical flow is shown in the regions in which there is greatest movement, identified as being the hands. Finally, image (f) illustrates the speaker diarization process with the two participants involved in the VOM session. Clusters belonging to each participant are obtained from the input signal, estimating the speech time of each segment, as well as the speech pauses/interruptions.

source separately. They can then be combined so as to define and recognize communicative indicators.

A. Related work

The Restorative Justice approach focuses on the personal needs of victims. As discussed above, achieving success in the VOM sessions depends largely on how the participants communicate with each other. A large number of techniques can be found in the literature for application in VOM. A good example of this is provided by Umbreit's handbook [citeumbreit:vom](#). This resource offers an empirically grounded, state-of-the-art analysis of the application and impact of VOM. It provides practical guidance and resources for VOM in the case of property crimes, minor assaults, and, more recently, crimes of severe violence, where family members of murder victims request a meeting with the offender. Since most of these cases addressed are of a highly sensitive nature, participants are likely to manifest emotional states, when interacting with the others, that can be physically observed through their non-verbal communication [5].

Recently, a number of studies have proposed ways in which personality traits can be inferred from multimedia data [6] and which can be applied directly to the approach taken by Restorative Justice. The prediction of these responses takes a

particular interest in meetings involving a limited number of participants. For instance, in [7] the goal was both to detect the social signals produced in small group interactions and to emphasize their importance markers. In addition, the works of [8], [9] combined several methodologies to analyze non-verbal behavior automatically by extracting communicative cues from both simulated and real scenarios. Thus, most of these social signal processing frameworks involve the detection of a set of visual indicators from the analysis of the participants' body and face. Additionally, information obtained from speech is commonly used [10], [11], [12], as is other information obtained from ambient and wearable sensors [13].

Many of the aforementioned studies demonstrate that indicators of agreement during communication are highly dependent on social signals. As such, it is possible to perform an exhaustive analysis to detect the role played by each participant in terms of influence, dominance, or submission. For instance, in [14], both the interest of observers and the dominant participants are predicted solely on the basis of behavioral motion information when looking at face-to-face (also called *vis-a-vis* or dyadic) interactions. Furthermore, there are many interdisciplinary, state-of-the-art studies examining related fields from the point of view of social computing, some of which are summarized in [3], [4].

In most of these frameworks, it can be observed that both ambient intelligence and egocentric computing methods are defined. Ambient intelligence refers to electronic environments that are sensitive and responsive to the presence of people, whereas egocentric computing refers to the use of wearable devices. Often, existing techniques of data acquisition make use of interface devices [15], or special items such as gloves [16] to increase recognition accuracy. However, while these techniques give impressive results in simulated environments, their use becomes largely infeasible in real-case scenarios due to their invasiveness and the uncontrollable nature of the application context. Because of the need to avoid wearing intrusive egocentric devices, some other ambient sensors that provide multi-modal data might be considered. In [8], a custom developed system is applied in a real-case scenario for job interviews. The data acquisition procedure is performed using different types of camera, by setting them up in different positions and with different ranges for capturing visual and depth information. Similarly, scenes with non-invasive systems have been proposed in other studies, such as [17], which provides trajectory analyses from body movements and gestures. Furthermore, audio information has been analyzed in [18], with the objective of modeling descriptors for speech recognition. This can be useful information for measuring the levels of activity from speech cues, including detection of speech/non-speech, interruptions, pauses, or segments obtained from a speaker diarization process.

The participants involved in these conversational settings usually appear either sitting or with some parts of their body occluded. Therefore, from a computer vision point of view, it might be best just to focus only on the upper body regions. Then, visual feature extraction techniques can focus on the most significant sources of information coming from the region of interest, which might be the face or hands, for example. These regions provide discriminative behavioral information, or adaptors, which are movements, such as head scratching, indicative of attitude, anxiety level and self-confidence [19]; or beat gestures, which are small baton-like movements of the hands used to emphasize important parts of speech with respect to the larger discourse [20]. However, as explained in [8], [21], body posture is also found to be an important indicator of a person's emotional state. Additionally, another potential source of information is provided by facial expressions [10], [11], [22], [23].

In order to analyze these visual features automatically most approaches are based on classic computer vision techniques applied to RGB data. However, extracting discriminative information from standard image sequences is sometimes unreliable. In this sense, recent studies have included compact multi-modal devices which allow 3D partial information to be obtained from the scene. In [24], the authors proposed a system for real-time human pose recognition including depth information for each image pixel. In this case, information is obtained by means of a Kinect™ device, which estimates a depth map based on the inverse of time response of an infrared sensor sampling within the scene. This new source of information, which provides visual 2.5D features, has been recently exploited for creating new human pose descriptors by

combining different state-of-the-art RGB-depth features [25], as well as they are used in a large amount of Human Computer Interaction (HCI) applications [26].

Once body features are computed, behavioral indicators can be analyzed by studying their trajectories using pattern recognition approaches. Some of these methods in this context are based on dynamic programming techniques such as Dynamic Time Warping (DTW) [27] or involve statistical approaches, such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [28], [29], [30].

Once data from the environment have been acquired and processed to define a set of behavioral features, they serve as the basis for modelling a set of communication indicators. For instance, in [31], the authors outline a system for real-time tracking of the human body with the objective of interpreting human behavior. However, in the context of conversations we are particularly interested in behavioral traits belonging to social signals captured in the communication and in the interactions between the participants in the VOM sessions. In this sense, levels of agitation (or energy), activity, stress, and engagement are analyzed not only from their body movements, but also from their speech, facial expressions, or gaze directions, in order to predict behavioral responses.

In this paper, we propose a non-invasive ambient intelligence framework for the analysis of conversations applied to real-case scenarios of VOM processes. Due to the ethical constraints imposed by the legal system, to date no datasets of VOM sessions have been made publicly available. Here, for the first time, a novel VOM data set has been collected. In terms of feature extraction, we extract a set of multi-modal audio-RGB-depth features and behavioral indicators, which are then used to measure the degree of receptivity, agreement, and satisfaction using state-of-the-art machine learning approaches and the ground truth defined by the mediators in the VOM sessions. As a result, we find that our technology achieves a high correlation between the most relevant features obtained by the behavioral indicators and the information provided by the experts.

The rest of the paper is organized as follows. Section 2 presents the material acquired and used in this study. In section 3, the various system modules are described. Section 4 outlines the proposal setup and the experimental results. Finally, section 5 concludes the paper.

II. DATA COLLECTION

This section describes the data acquisition process, as well as how the annotations of social variables have been defined.

First, an environmental study was undertaken in the various rooms in which recording was to take place, and in which the non-invasive devices were to be set up.

Once the environmental study had been completed, decisions regarding the ethical constraints that had to be satisfied were taken in order to protect the recorded data. This procedure involved the drawing up of three fundamental ethical documents: the researchers' signed undertaking, informed consent, and the case-codification.

As the sessions typically involve two or three participants, the homogeneous distribution of the cameras enabled us to

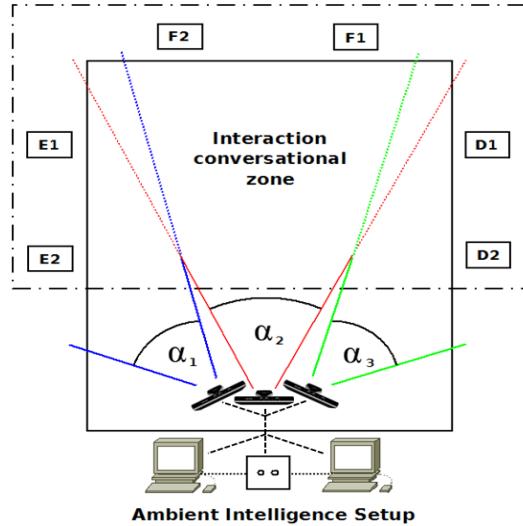


Fig. 2. Acquisition architecture. $E_1, E_2, F_1, F_2, D_1, D_2$ are the participants codified by their respective positions (E: left, F: front, D: right); the angles of view for the different cameras are the same, and hence $\alpha_1 = \alpha_2 = \alpha_3$.

capture at most two people-per-camera. Specifically, the devices used were three Kinect™ sensors and two laptops (which varied depending on the number of participants). Thus, a maximum of six people could be recorded¹. Figure 2 shows the ambient intelligence setup with all the elements involved and their distribution.

Recordings were made in various towns and cities of Catalonia. Figure 3 shows the distribution of these recording sites together with the number of sessions recorded in each. Most recordings were made in the capital city of Barcelona with a total of 15 sessions, followed by Vilanova i la Geltrú with a total of four. Two sessions were recorded in each of Manresa, Tarragona, and the youth penitentiary center in Granollers. Finally, one session was recorded in Terrassa.

Thus, 26 VOM sessions, of 35 minutes-per-session, were recorded (average duration ranging from 20 minutes to 2 hours depending on the session), in which a mediator engaged in a conversational process with different parties. Of the total number of sessions, 15% were joint encounters, with both parties (victim and offender) being present in the VOM. The remaining sessions were individual encounters involving one or other of the parties and the mediator. As indicated in the VOM program itself, the participants include both adult men and women, and they fulfill different participant roles: victim, offender, or mediator. Some of the sessions also involved accompanying persons, either a professional from the specific center, or experts in some particular field relevant to the case under discussion.

Each recorded session² provided audio-RGB-depth information. These modalities were registered using the camera param-



Fig. 3. Map with recorded regions marked with dots.

eters, and synchronized between the various devices through the system clock. The set of images for each session were recorded at a resolution of 640×480 and at an average of 12 frames per second (fps), both for RGB and depth information. Each audio channel, belonging to one of the four microphones spread out linearly along a multi-array microphone, processed 16-bit audio at a sampling rate of 16 kHz. The distance between participants and the Kinect™ device was between 1 and 2 meters depending on the specific characteristics of the recording facility.

As the data protection regulations only allow one mediator to annotate each session, the annotators were those mediators that had greatest familiarity with the case being dealt with in each session. Only in a few isolated cases were there two mediators in the session. Thus, in some cases the questionnaires completed by the mediators, recording their impressions and feelings regarding the party/ies and the overall sessions, were subsequently confirmed by a second mediator from the team so as to guarantee the consistency of the defined ground truth values. The system responses were determined by considering both the state-of-the-art methods for the study of behavioral traits in people involved in similar scenarios, as presented in section I-A [2], [3], [4], [6], [7], [8], [9], [10], [11], [12], [13], [14], and in the subsequent discussion held with the mediators, taking into account the aims of their work with the Department of Justice. Finally, we defined the system's ground truth as:

- **Receptivity:** degree of engagement shown by each party during the session.
- **Agreement:** degree of agreement reached between the parties (quantified globally for each session).
- **Satisfaction:** degree of agreement reached between the parties in relation to the mediator's expectations (quantified globally for each session).

These terms do not only depend solely on the individual's personality, but they are also dependent on the interaction with other participants, and hence they are a consequence of putting

¹The maximum number of people in the recorded sessions was five.

²See an example of the different modalities and visual extracted visual features in the **supplementary video material sample**.

the behavioral traits into practice. The quantitative nature of these social responses was validated by a randomly selected mediator who had not been involved in that case so as to obtain a more objective evaluation. This approach was likewise applied to two features describing the evolution in the level of nervousness manifest by each party at the beginning and at the end of the process, respectively. Therefore, for each session and for each party, mediators ranked the observed quantity of these behavioral indicators from 1 to 5, where 1 is the lowest value and 5 the highest. Table I shows a numerical summary of the data acquired.

TABLE I. SUMMARY OF DATA ACQUIRED.

Individual encounters	22
Joint encounters	4
Total sessions	26
Penitentiary centers	1
Office centers	4
Total justice centers	5
Mediators	7
Parties	30
Total n° participants	37
Total n° frames	1,436,400
Average n° minutes/session	35

III. PROPOSED METHOD

The method proposed for the non-verbal communicative analysis is described in this section. The framework consists of three main sequential modules illustrated in Figure 4. The first module includes the multi-modal feature extraction from audio-RGB-depth data, which is described in Figure 5. As shown in the scheme, the steps for obtaining multi-modal features from different sources of information are the speaker diarization, user segmentation, and region detection. Once the multi-modal features have been extracted, they are used to define the behavioral indicators in the second module, which

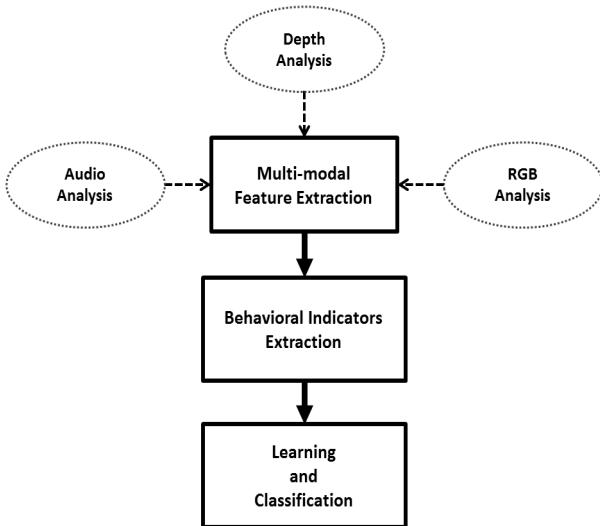


Fig. 4. Modules of the proposed system.

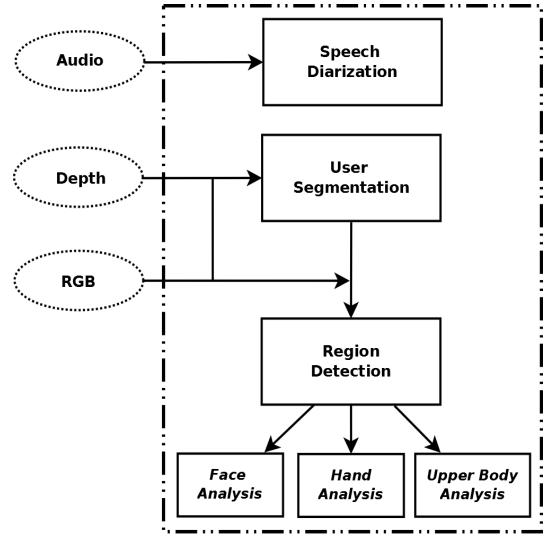


Fig. 5. Multi-modal feature extraction module.

are used as inputs for the learning and classification steps in the system's third module.

For all the system's modules, let $V = \{v_1, v_2, \dots, v_r\}$ be a set of r recorded VOM sessions and $C = \{c_1, c_2, \dots, c_s\}$ be a set of s VOM cases. Since a case is divided into one or more VOM sessions, one session $v \in V$ may belong either to the same case as another session, or to a different case. Thus, a subset $\{v_1, v_2\} \in V$ belonging to the first case c_1 is then denoted as $\{v_1^1, v_2^1\}$, where the super-indices indicate the number of the case.

The rest of this section describing multi-modal feature extraction is structured as shown in Figure 5, where each main block is described in a different section and where it is dependent on both the information modality and the previous block. Thus, following the remaining modules of Figure 4, we describe the behavioral indicators and finally the learning and classification of receptivity, agreement, and satisfaction labels.

A. Audio Analysis

This section describes the feature extraction applied to the audio data source. As audio cues are the primary source of information, they can provide useful evidence for speech production. Although in noisier environments these cues do not easily distinguish between the user that is speaking from the other participants (as a result of overlapping speech), in VOM scenarios this is not such a problem given the small number of participants.

1) Speaker Diarization: In order to obtain the audio features, we use a diarization scheme based on the approach presented in [32]. These features correspond to state-of-the-art methods for audio descriptions, which have been successfully applied in several audio analysis applications [33], [34], [35]. The process is described below:

Description: The input audio is analyzed using a sliding-window of 25 ms, with an overlap of 10 ms between consecutive windows, and each window is processed using a short-time

Discrete Fourier Transform (DFT), mapping all frequencies to the Mel scale. A more precise approximation of this scaling for frequencies used in Mel Frequency Cepstral Coefficients (MFCC) implementations, is represented as:

$$\hat{f}_{mel} = k_{const} \cdot \log_n \left(1 + \frac{\hat{f}_{lin}}{F_b} \right), \quad (1)$$

where F_b and k_{const} are constant values for frequency and scale, respectively. The Koenig scale \hat{f}_{lin} is exactly linear below 1000 Hz and logarithmic above 1000 Hz. In brief, given N -point DFT of the discrete input signal $\tilde{x}(n)$,

$$\tilde{X}(k) = \sum_{n=0}^{N-1} \tilde{x}(n) \cdot \exp \left(-\frac{j2\pi nk}{N} \right), \quad k = 0, 1, \dots, N-1, \quad (2)$$

a filter bank with several equal height triangular filters is constructed. Each of these filters has boundary points expressed in terms of position, which depends on the sampling frequency and the number of points N in the DFT. Finally, the Discrete Cosine Transform (DCT) is used to obtain the first 13 MFCC coefficients. These coefficients are complemented with the first and second time-derivatives of the Cepstral coefficients.

Speaker segmentation: Once the audio data are properly described by means of the aforementioned features, the next step involves identifying the segments of the audio source which correspond to each speaker. A first coarse segmentation is generated according to a Generalized Likelihood Ratio, computed over two consecutive windows of 2.5 s. Each block is represented using a Gaussian distribution, with a full covariance matrix, over the extracted features. This process produces an over-segmentation of the audio data into small homogeneous blocks. Then, a hierarchical clustering is applied to the segments. We use an agglomerative strategy, where initially each segment is considered as a cluster, and at each iteration the two most similar clusters are merged, until the stopping criterion of the Bayesian Information Criterion (BIC) is met. As in the previous step, each cluster is modeled by means of a Gaussian distribution with a full covariance matrix and the centroid distance is used as the link similarity. Finally, a Viterbi decoding is performed in order to adjust the segment boundaries. Clusters are modeled by means of a one-state HMM using GMM as our observation model with diagonal covariance matrices. Figure 1 (f) represents an example of this procedure, showing the clusters where the speech signal falls at each instant. Since most of the participants appear in just a single mediation session, we do not learn any speaker models from the cluster GMMs. Therefore, models extracted from one session are not used in the diarization process of other sessions.

B. User Detection

Both RGB and depth data are used for the postural and behavioral analyses of the parties (for examples of these images see Figure 1 (a) and (b), respectively). In this sense, the first step involves performing a limb-segmentation of the body based on the Random Forest method of [24]. This process is

performed by computing random offsets of depth features as:

$$f_\theta(I, \dot{p}) = d_I \left(\dot{p} + \frac{\mu}{d_I(\dot{p})} \right) - d_I \left(\dot{p} + \frac{\nu}{d_I(\dot{p})} \right), \quad (3)$$

where $d_I(\dot{p})$ is the depth at pixel \dot{p} in image I , and parameters $\theta = (\nu, \mu)$ describe offsets ν and μ . Since offsets are normalized by $\frac{1}{d_I(\dot{p})}$, features are 3D translation invariant. Each split node consists of a feature f_θ and a threshold τ . To classify pixel \dot{p} in image I , one starts at the root and repeatedly evaluates Eq. 3, branching left or right according to the comparison to threshold τ . At the leaf node reached in tree t of the Random Forest, a learned distribution $P_t(l|I, \dot{p})$ over body part labels l is stored. The distributions are averaged together for all the trees in the forest to obtain the final classification:

$$P(l|I, \dot{p}) = \frac{1}{T} \sum_{t=1}^T P_t(l|I, \dot{p}). \quad (4)$$

Then, we choose body part labels with high probability values so as to detect those pixels belonging to the person. Figure 1 (c) shows a user detection example of applying this segmentation.

Once regions of interest have been located, it is of particular interest to obtain real-world distance values for certain computed features so that they are comparable between different subjects. To do this, we employed a similar procedure to that explained in [36], which converts the 2D pixels into 3D real-world coordinates using the Kinect™ depth values. However, since these raw sensor values returned by the depth sensor are not directly proportional to the depth, in [36], they scale with the inverse of the depth. Therefore, each pixel (\dot{x}, \dot{y}) of the depth camera can be projected to metric 3D space as:

$$x = (\dot{x} - \delta_x) \frac{d(\dot{x}, \dot{y})}{\kappa_x}, \quad y = (\dot{y} - \delta_y) \frac{d(\dot{x}, \dot{y})}{\kappa_y}, \quad z = d(\dot{x}, \dot{y}), \quad (5)$$

where (x, y, z) will be the real world coordinates, and δ_x , δ_y , κ_x , κ_y , the intrinsics of the depth camera. These values will be computed over the detected interest regions in order to define the communicative indicators described in next sections.

C. Region Detection

This section describes the different feature extraction modules applied to the visual data source once the user has been segmented. Specifically, we perform an analysis of the face, hands, and upper body, as well as visual movements in these regions during conversations. Moreover, the estimation of head and body postures provides information about where each person in the VOM sessions directs their gaze, and the status of the people in terms of agitation, respectively. Given that it is not feasible to annotate manually all the regions in several thousands of frames to measure the recognition rates of the detection methods used in this section, we propose a heuristic procedure which maintains the continuity of region detections, removing false positives and recovering false negative regions, considerably reducing the manual interaction rate.

1) *Face Analysis*: We are primarily concerned with obtaining the head pose angle of each of the participants in the session. To do this, we base our approach on that of [37] which uses a set of face models. The face model is based on a mixture of trees with a shared pool of parts B , where every facial landmark is modelled as a part and global mixtures are used to capture topological changes due to viewpoint. Global mixtures can also be used to capture gross deformation changes for a single viewpoint, such as changes in expression. Each tree is written as $T_m = (B_m, E_m)$ as a linearly-parameterized, tree-structured pictorial structure [38], where m indicates a mixture and $B_m \subseteq B$. Let I be an image, and $\rho_i = (x_i, y_i)$ the pixel location of part i . Configuration of parts $L = \{\rho_i : i \in B\}$ are then scored as:

$$S(I, L, m) = App_m(I, L) + Shape_m(L) + \alpha^m, \quad (6)$$

$$App_m(I, L) = \sum_{i \in B_m} w_i^m \cdot \phi(I, \rho_i), \quad (7)$$

$$Shape_m(L) = \sum_{ij \in E_m} \gamma_{ij}^m dx^2 + \eta_{ij}^m dx + \varrho_{ij}^m dy^2 + \varphi_{ij}^m dy. \quad (8)$$

Eq. 7 sums the appearance evidence for placing a template w_i^m for part i , tuned for mixture m , at location ρ_i . The feature vector $\phi(I, \rho_i)$ (e.g., HoG descriptor) is extracted from pixel location ρ_i in image I . Eq. 8 scores the mixture-specific spatial arrangement of parts L , where $dx = x_i - x_j$ and $dy = y_i - y_j$ are the displacement of the i th part relative to the j th part. Each term in the sum can be interpreted as a spring that introduces spatial constraints between a pair of parts, where the parameters $(\gamma, \eta, \varrho, \varphi)$ specify the rest location and rigidity of each spring. Finally, the last term α^m is a scalar bias or "prior" associated with viewpoint mixture m . Eq. 6 requires a separate template w_i^m for each mixture/viewpoint m of part i . However, parts may look consistent across some changes in viewpoint. In the extreme cases, a "fully shared" model would use a single template for a particular part across all viewpoints, $w_i^m = w_i$. A continuum between these two extremes is explored, written as $w_i^{f(m)}$, where $f(m)$ is a function that maps a mixture index (from 1 to M) to a smaller template index (from 1 to M'). Various values of M' are explored: no sharing ($M' = M$), sharing across neighboring views, and sharing across all views ($M' = 1$).

The detection of the head pose angle is performed by averaging HOG feature as a polar histogram over 18 gradient orientation channels, as computed from the entire PASCAL 2010 data set [39]. In Figure 1 (d) we can visualize the set of computed features plotted on the detected face.

While face detection takes place for each tested image, a semi-automatic heuristic procedure is proposed in order to improve the continuity of positive detections of regions of interest in the person between consecutive frames, and to correct possible erroneous detections due to the inherent difficulties of the problem at hand. Figure 6 shows the flowchart of the procedure applied to each frame. In short, it consists of a temporal filtering methodology of detected regions (faces) between one-by-one consecutive frames. It is

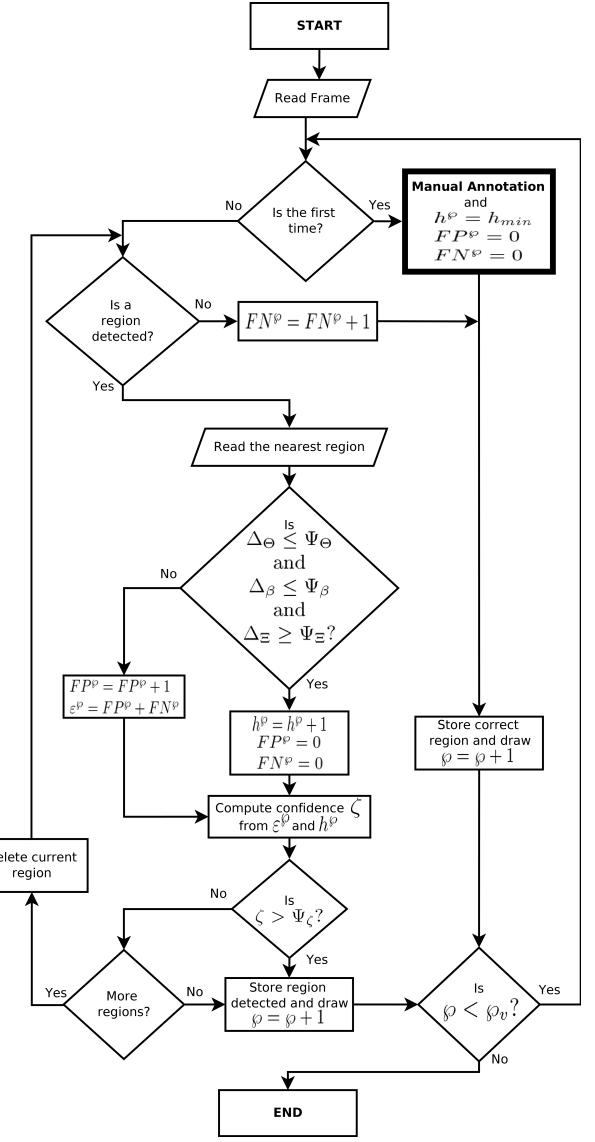


Fig. 6. Flowchart of the heuristic procedure applied to each frame. The total number of people that appear in the current video v is denoted by φ_v . Constraints of the main condition at the center of the flowchart are denoted by Δ_Θ , Δ_β , Δ_Ξ , and their respective thresholds Ψ_Θ , Ψ_β , Ψ_Ξ . The counting variables are FN^v , FP^v , h^v , representing the accumulated number of false negatives, false positives, and hits for the current person v . They are used to compute the confidence ζ from the accumulated detection errors ε^v and the hits h^v , and to decide whether the current detected region has to be stored or discarded through the threshold Ψ_ζ .

based on three main constraints that enable us to choose the detected regions in the current frame by comparison to the previous one: offset pixels produced by the mass centers, offset angle produced by head poses, and the size difference factor produced among the region areas. Thus, three thresholds Ψ_Θ , Ψ_β , and Ψ_Ξ are respectively used to discriminate the occurred cases on each constraint, whose values may vary depending on the session conditions. Moreover, there are three counting variables that accumulate, for each person, the number of

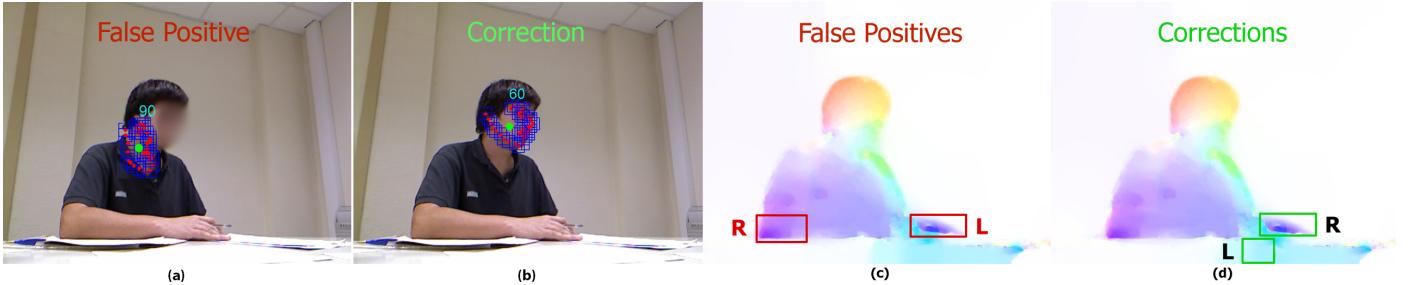


Fig. 7. Examples of how the semi-automatic heuristic procedure works on two pairs of frames in different sessions. The correction of false positives and false negatives is shown, improving the continuity of the detection of positive regions of interest between consecutive frames. Image (a) shows a false positive detection for the face region, whereas image (b) shows its correction with the proper fitting. Image (c) shows false positive detections for the hand regions, choosing those blobs obtained by means of skin segmentation having highest optical flow with respect to the previous frame. Image (d) shows the correction of these regions by comparing them with the positive hand detections from the previous frame.

correct detections (hits) h^φ , false positives FP^φ and false negatives FN^φ . Then, a confidence ζ is computed from h^φ and the sum of false detections ε^φ to decide whether the current detected region has to be stored or discarded by means of the threshold Ψ_ζ . These counting variables are highly dependent on constraint thresholds, as they make the system more or less restrictive when choosing detected regions. Therefore, a trade-off between constraint thresholds and control thresholds should be reached when assigning their values in order to assure the continuity of positive region of interest detections for that person (even though the method could not detect any region in the image), and to decide whether a manual annotation is required to re-initialize the detection process in the (approximately) desired frequency rate. Images (a) and (b) of Figure 7 show an example of how the heuristic procedure corrects a false positive detection on the face region for a given frame. In the end, so that all image sequences were annotated with the head pose analysis, we found that the use of the proposed heuristic reduced manual interactions by a ratio of up to five.

2) Hand Analysis: Given that the skeletal model computed from the person segmentation image [24] does not offer an accurate fit of the hand joints in our particular scenario, we designed a semi-automatic procedure for hand detection.

First, hands are manually annotated in the starting frames of each session to perform posterior color segmentation for the rest of the frames. In this way, a GMM is learned with the marked set of most significant pixels, defining the skin color model of the person. Then, subsequent frames are tested within the GMM built using a threshold ϑ , discriminating those pixels belonging to the skin color from those belonging to the background. The resulting blobs are filtered using mathematical morphology closing operation with a 3×3 square structured element to discard noise and to obtain smoother regions. Once the set of blobs has been obtained, we need to choose those two candidates that belong to the hand regions. This is performed by computing the optical flow between consecutive frames, which allows to discard noise in those cases in which we obtain more than two blobs by retaining those with higher movement. Figure 1 (e) illustrates this procedure.

On the other hand, we use the same heuristic procedure

as that applied to the face analysis step for choosing the two best hand candidates. Images (c) and (d) of Figure 7 show an example of how the heuristic procedure corrects false positive detections on the regions of the hands. The incorrect regions detected in the first instance are the blobs presenting the highest optical flow, and then the heuristic procedure corrects these regions by comparing them with the hand regions obtained from the previous frame. As in face detection, manual annotation may be required in those cases where the heuristic procedure needs to be re-initialized. For this task, an interface has been designed for the manual annotation of the hand regions for the set of frames in which this occurs. When the user makes any annotation, the GMM color model is newly re-constructed at this frame using the marked pixel positions, and the whole process is repeated. In this case, using the proposed heuristic we also found similar reduction regarding manual interaction effort as in the case of face region detection.

Once we have obtained the blobs belonging to the hand regions, the extremes with higher optical flow magnitude are used to obtain 2D hand positions. Finally, these positions are transformed to 3D real world coordinates using the conversion of Eq. 5.

3) Upper Body Analysis: As presented above, the probability of each pixel of an image belonging to a labeled body part is computed in Eq. 4 using depth features of Eq. 3. This information is used for the subsequent calculation of optical flow on RGB images where the upper body region appears. Therefore, each pixel p of the image I detected by Random Forest with high probability of being part l of the person is used to calculate the optical flow. Finally, an average $\bar{\sigma}$ of optical flow is computed for the upper body region. An example of user detection where upper body region is highlighted is shown in Figure 1 (c). Average optical flow $\bar{\sigma}$ is later used to define behavioral indicators.

D. Behavioral Indicators

Once the multi-modal features have been extracted, we can use them to build a set of behavioral indicators that reveal communicative cues about each party involved in the VOM process. This information is of great interest in detecting the

response of subjects to certain feelings or emotional states during the conversation [5]. In particular, since the behavioral cues of the mediators are not of interest for our purposes here, we focus mainly on those of the parties.

Thus, in this section, we describe the set of behavioral indicators constructed from the output of the different blocks shown in Figure 5, which define the final feature vector for each party within the VOM process.

1) Target Gaze Codification: The head pose and the face is obtained by applying the methodology explained in section III-C1. Therefore, in a given session, we compute the correlations between the head pose angles belonging to each participant and the positions taken by the rest of the participants in that session. This procedure is performed in order to identify the visual focus of attention among the different participants in the conversation [9]. For this purpose, different ranges are assigned to each participant in terms of angle limits. Given that the participants belonging to the same party are seated in adjacent positions (see acquisition architecture in Figure 2), each range represents a possible participant vision field of his/her gaze towards the target party. Thus, given a frame of the session and a participant, if his/her head pose angle falls within a particular range, then the party found within that range is identified as the target gaze of this participant for that frame, which means the participant is looking at this party. Since sessions have different setups, they may consist of one or two parties (and the mediator), each with a different number of participants. Therefore, the ranges require manual assignment depending on each session setup. Then, the target gazes are automatically identified for all the frames of the session. Table II describes all the target gaze combinations that can be found in any possible session setup. Note that we codify the target gazes with values 1 or 0 depending on whether the person is looking at the party or not, respectively. It is also important to note that we focus on the target gazes between the parties, instead of the participants. Thus, the set of target gazes towards the parties is codified as:

- P: Looking at the same party, whether it is the offender (Off) or the victim (Vic). If there is more than one party (case of a joint encounter), P changes on the mediator column by 'Off' or 'Vic' depending on the party.
- M: This party is looking at the mediator.
- MP: This party is looking at the same party.
- MO: This party is looking at the other party.

TABLE II. ALL POSSIBLE COMBINATIONS FOR THE CODIFICATION OF TARGET GAZES.

	Mediator	Part
1 party with only 1 person	P 0 0	0 M 0
1 party with several people	P 0 0	MP M 0
2 parties with only 1 person in one party	Off 0 Vic	0 M MO
2 parties with several people in one party	Off 0 Vic	MP M MO

Figure 8 (a) shows an example of crossed gazes between the mediator and a party in a real VOM session. Finally, we use these binary values to compute the time percentages of target gazes for each party. Therefore, for any given party, there exists a total of 4 indicators for representing all possible combinations of target gazes.

2) Agitation Estimation: As explained in section III-C2, 3D positions belonging to the hand regions are computed from the extreme positions of higher optical flow. From these positions, we are able to quantify the movement for each region between consecutive frames. For this purpose, let $F = \{\iota_1, \iota_2, \iota_3, \dots, \iota_\lambda\}$ be a set of consecutive frames. This set of frames belongs to a video session $v \in V$, being $\lambda = r$ the maximum length of the set. Then, for each region we compute the average agitation over all the frames $\iota \in F$ as:

$$A_h = \frac{1}{\lambda} \sum_{\iota=1}^{\lambda} \Delta_h^\iota, \quad (9)$$

where $\Delta_h^\iota = \Delta_p^\iota + \Delta_q^\iota$ are the displacements among 3D positions of hands h (left p and right q) between frames ι and $\iota - 1$, computed using Euclidean distance. Therefore, A_h contains the accumulated average of displacements produced by both hands between frames F .

On the other hand, in section III-C3 we explained how the average optical flow $\bar{\sigma}$ is obtained for the upper body region. Therefore, if we denote as $\bar{\sigma}_\iota$ the average optical flow of the upper body for a given frame $\iota \in F$, then:

$$A_b = \frac{1}{\lambda} \sum_{\iota=1}^{\lambda} \bar{\sigma}_\iota, \quad (10)$$

where A_b contains the accumulated average of optical flow produced by the upper body between frames F .

In short, for each party and session, agitation averages are computed over processed frames, with a total of 8 agitation indicators, both alone and combined with the other indicators previously calculated. The idea of combining these indicators with other behavioral features is inspired by [40], [14]. In this case, we consider a combination between the features describing the agitation from the upper body and those describing the hands while looking at the participants (see list of Table III).

3) Posture Identification: From the 3D body position, we detect the body posture as 1 behavioral indicator, which may describe the engagement (or involvement) of the party within the VOM session. Our description of body posture is classified into three main positions (tilted backward, normal, tilted forward), where the posture selected is the one that has the most occurrences over the processed frames.

In addition, 3D hand positions are used to detect where the hands are along the processed frames, in terms of average and time percentages. In particular, we discriminate three cases (i.e., 3 indicators): hands together, hands touch the face, and hands under the table. This is done in a similar way as for the agitation estimation, using Euclidean distance computed over 3D positions.

- Hands together: We compute for each frame the distance between left and right hand positions belonging to the target subject, and we consider the frames where the distance values are below that of the threshold Ψ_Θ . Finally, we compute the time percentage for those frames where the target subject appears with their hands together.
- Hands touch the face: We compute for each frame the distance between each hand position and the position

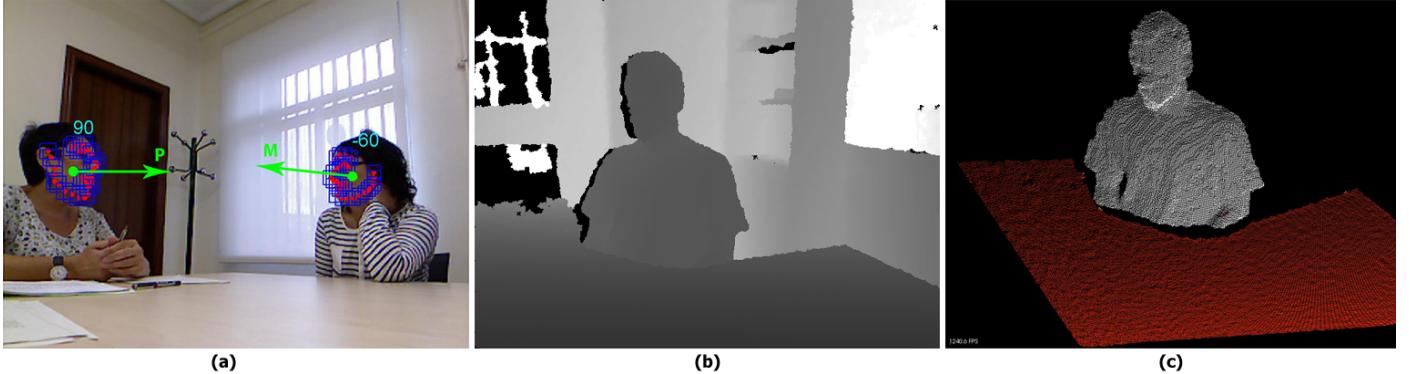


Fig. 8. Visual instances of some situations where behavioral indicators are detected in VOM sessions. Image (a) shows the detection of crossed gazes between the mediator and the other participant. Images (b) and (c) show a depth image and its segmentation for the person (white point cloud) and the table (red point cloud), respectively, which is used to detect a situation in which the target subject appears with his or her hands under the table.

belonging to the face center of mass obtained in section III-C1. Then, we consider the frames where the distance values are below that of the threshold Ψ_{Θ} . Finally, we compute the average distance for those frames where the target subject appears with their hands touching their face.

- **Hands under the table:** For each frame, we first perform a segmentation of the tables using [41] to obtain planar objects within images. Then, we compare the 3D positions of both hands with the position of the tables in order to discriminate the two possibilities where the hands may appear under or above the table. Finally, we compute the time percentage for those frames where the target subject appears with their hands under the table. Figure 8 (b) and (c) illustrate an example of this procedure, showing respectively the input depth image and its segmentation.

4) Speech Turns/Interruptions Detection: The speaker diarization process of section III-A1 detects time segments belonging to each participant in the VOM process. In order to extract the degree of interaction, we not only use the length of time during which each participant speaks, but we also count the number of turns in each session. This enables to differentiate between a session where each party expresses its position from a session in which a conversation is maintained between the VOM participants. Apart from the quantification of turn taking, a relevant indication in the social communication analysis is the detection of interruptions, which are related to the dominance and respect between two persons [42]. Using the time between turns, we compute the percentage of turns in which a participant interrupts another one. For instance, in the first three turns of Figure 1 (f) a participant (red) interrupts the mediator (green), while the mediator waits until the other participant ends his turn before starting to speaking again.

E. Classification

The total number of behavioral indicators is 36, which define the feature vector for each sample in our data set. Here, we define a sample as each party participating in a VOM session $v \in V$. Thus, if a session v involves two parties and the

mediator, we introduce one sample of 36 features for each of the two parties. On the other hand, if a session involves just one party and the mediator, we introduce only one sample corresponding to the party involved. Table III summarizes the behavioral indicators and gives a brief description.

TABLE III. SUMMARY OF BEHAVIORAL INDICATORS DEFINING EACH FEATURE VECTOR. THE LAST TWO FEATURES f_{35} AND f_{36} ARE OBTAINED FROM SURVEY ANSWERS PROVIDED BY MEDIATORS.

Feature	Brief description
f_1	Party's role within the VOM session (victim or offender)
f_2	This party looks at the other
f_3	The other party looks at this party
f_4	This party looks at the mediator
f_5	The mediator looks at this party
f_6	Body posture inclination of this party
f_7	Gender of the mediator
f_8	Gender of this party
f_9	Gender of the other party
f_{10}	Age of the mediator
f_{11}	Age of this party
f_{12}	Age of the other party
f_{13}	Session type (individual/joint encounter)
f_{14}	Upper body agitation of this party
f_{15}	Upper body agitation of this party while looking at the other party
f_{16}	Upper body agitation of this party while looking at the mediator
f_{17}	Hands agitation of this party
f_{18}	Hands agitation of this party while looking at the other party
f_{19}	Hands agitation of this party while looking at the mediator
f_{20}	Hands agitation of the mediator while looking at this party
f_{21}	Hands agitation of the other party while looking at this party
f_{22}	This party has the hands together
f_{23}	Hands of this party touches his/her face
f_{24}	This party has the hands under the table
f_{25}	Mediator speaking time
f_{26}	Speaking time of this party
f_{27}	Speaking time of the other party
f_{28}	Mediator speaking turns
f_{29}	Speaking turns of this party
f_{30}	Speaking turns of the other party
f_{31}	Mediator interrupts this party
f_{32}	This party interrupts the mediator
f_{33}	This party interrupts the other party
f_{34}	The other party interrupts this party
f_{35}	Nervousness of this party at the beginning
f_{36}	Nervousness of this party at the end

As explained in section II, the observations of the classification task are the accuracies achieved by the system when

predicting receptivity, agreement, and satisfaction. Then, the correlation can be observed between the observations predicted by the system and the impressions recorded by the mediators. These opinions are quantified values of receptivity, agreement, and satisfaction presented in relation to the parties involved in the VOM session, and represent the ground truth of our system. The ground truth values are assigned to each sample of the data set. Since agreement and satisfaction are globally assigned for each session, those sessions containing two parties will share the same ground truth labels of agreement and satisfaction for both generated samples, meanwhile the receptivity ground truth value is assigned to each sample (party) independently.

Learning is then performed on these samples and their features as a binary classification problem, grouping into two classes the quantifications performed by the mediators. To do this, we employ four classical techniques from the machine learning field: AdaBoost [43], Support Vector Machines (SVM) using a Radial Basis Function (RBF) [44], Linear Discriminant Analysis (LDA), and three kinds of Artificial Neural Networks (ANN), in particular Probabilistic Neural Networks (PNN) [45], and Cascade-Forward (CF) and Feed-Forward neural networks (FF) [46]. In addition to the binary classification analysis we also conduct a regression study using epsilon-SVR (Support Vector Regression) [44] in order to predict continuous quantifications of the three labels.

IV. EXPERIMENTS

This section describes the experiments performed when using the behavioral indicators summarized in Table III. First, we describe the setting and validation measurements, before outlining the experiments performed.

A. Setting and validation measurements

The parameters used in the heuristic procedure were experimentally set to ranges $\Psi_\Theta \in [50, 120]$, $\Psi_\beta \in [30, 60]$, and $\Psi_\Xi \in [0.1, 0.3]$, depending on the session, and the standard value $\Psi_\zeta = 0.5$ for all sessions.

The measurements for the features $\{f_2, f_3, f_4, f_5, f_{22}, f_{24}, f_{25}, f_{26}, f_{27}\}$ are time percentages, whereas the features $\{f_{14}, f_{15}, f_{16}, f_{17}, f_{18}, f_{19}, f_{20}, f_{21}, f_{23}\}$ contain averaged values of optical flow or distances, both taking into account the processed frames of a session. Features $\{f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}\}$ are turn taking percentages, where a turn means that the speaker changes.

Finally, the remaining features are codified into discrete values, either quantified to ranges $[1, 2]$ or $[1, 3]$ integer values for features $\{f_1, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}\}$, or to range $[1, 5]$ integer values for the nervousness features $\{f_{35}, f_{36}\}$. Each party of a video session v is a sample for the classification task, and the total number of used samples is 28.

In addition, an alternative was implemented where some features are divided into two -one belonging to the first half, the other to the second half of the session-. This procedure was initially performed to identify behavioral changes in subjects during the different halves of the session. However, no significant differences were found and, hence, we finally used the set of features without any temporal segmentation.

TABLE IV. ACCURACY CONSIDERING THE FIRST GROUPING CASE AND ALL FEATURES.

Label	AdaBoost	LDA	PNN	CF	FF	SVM
Satisfaction	57%	32%	57%	57%	86%	57%
Agreement	50%	54%	64%	64%	75%	64%
Receptivity	64%	50%	71%	71%	68%	75%

TABLE V. ACCURACY CONSIDERING THE SECOND GROUPING CASE AND ALL FEATURES.

Label	AdaBoost	LDA	PNN	CF	FF	SVM
Satisfaction	82%	43%	21%	82%	75%	82%
Agreement	71%	43%	29%	71%	75%	75%
Receptivity	75%	36%	39%	68%	75%	61%

Learning is performed using leave-one-out validation, keeping one sample out of the testing each time. Since the total number of samples is small and the ground truth values are quantified within ranges $[1, 5]$ (as for the nervousness features), we simplified the problem by grouping the different response degrees into binary groups, but we also performed a posterior regression analysis. In the case of a binary setup, the value 3 can be considered as being either high or low. For this reason, we ran the experiments twice to test each consideration.

In our experiments, we awarded the standard value of 50 to the number of decision stumps in the AdaBoost technique. For the SVM-RBF and epsilon-SVR, we experimentally set the cost, gamma, and epsilon parameters by means of the leave-one-out validation for each social response and minimizing regression deviation on the training set. Finally, we applied the same tuning procedure for the three standard neural network parameters: a Probabilistic Neural Network (PNN) with a spread value of 0.1 for the radial basis functions, and Cascade-Forward (CF) and Feed-Forward (FF) neural networks, both with a single hidden layer with 10 neurons values and Levenberg-Marquardt back-propagation training function. The results obtained are shown in terms of accuracy percentages.

B. Results and discussion

Due to the sensitive nature of the VOM process, never before (to the best of our knowledge) have mediators recorded their sessions so that they might subsequently analyze the cases. In this respect, therefore, the first results to emerge from this study are the session videos themselves, which are valuable materials via which the mediators can share their experiences and obtain feedback to improve their mediation skills.

As described in section III-C1, user manual interactions are a key requirement in our proposed semi-automatic system for improving the continuity of detections of positive regions of interest between consecutive frames. For our sessions, the average frequency rate of manual annotations required is 1 for each 2000 frames based on the above parameters. This means that when using these parameters the feature extraction procedures for both hands and faces offer high performance.

The predictions addressed in our classification task focus on three indicators: the degree of receptivity of the parties, the level of agreement reached, and the degree of mediator satisfaction. Tables IV and V show the results obtained when employing the different techniques and using the complete set

TABLE VI. ACCURACY CONSIDERING THE FIRST GROUPING CASE AND WITHHOLDING THE NERVOUSNESS FEATURES.

Label	AdaBoost	LDA	PNN	CF	FF	SVM
Satisfaction	57%	57%	57%	68%	64%	57%
Agreement	50%	43%	64%	57%	71%	64%
Receptivity	68%	46%	71%	75%	68%	75%

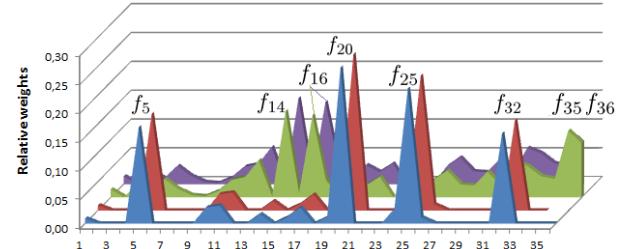
TABLE VII. ACCURACY CONSIDERING THE SECOND GROUPING CASE AND WITHHOLDING NERVOUSNESS FEATURES.

Label	AdaBoost	LDA	PNN	CF	FF	SVM
Satisfaction	82%	61%	21%	71%	86%	82%
Agreement	71%	57%	29%	71%	79%	75%
Receptivity	79%	46%	39%	64%	71%	61%

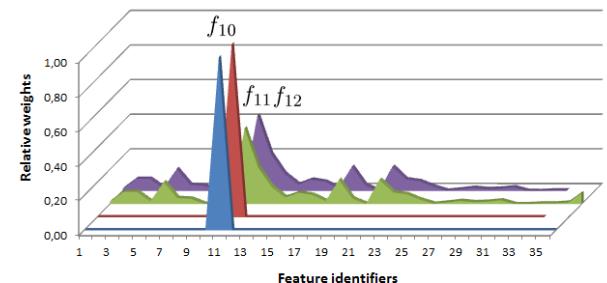
of behavioral indicators in Table III. In the first case (table IV), the degrees of quantification are grouped to obtain the binary setup as $\{1, 2, 3\}$ on the one hand, and $\{4, 5\}$ on the other. This hypothesis is also applied to the second case V, where the degrees quantified are grouped as $\{1, 2, 3\}$ versus $\{4, 5\}$. The most accurate results for each classifier are shown in bold. Note that as the features of nervousness are subjective indicators that are not automatically computed, we repeated the experiments without these two features $\{f_{35}, f_{36}\}$. These results are shown in Tables VI and VII, where the prediction is also analyzed under the grouping hypotheses. Once again, the results show a correlation between the features extracted and the categories selected. The percentage degree of accuracy in the predictions is then compared for the different techniques: AdaBoost, SVM, LDA, PNN, CF, and FF. It can be noted that, except for PNN and LDA (which are not good techniques for use with our dataset), all the classifiers are able to make predictions about the random decision. This indicates that there is a correlation between the captured data and the information that we want to predict. The most predictable social response is that of satisfaction, presenting an accuracy of 86% with the FF, followed by 82% with AdaBoost, SVM, and CF. The best result when predicting agreement was an accuracy of 79% with FF and, similarly, when predicting receptivity, the best accuracy was 79% with AdaBoost. These results are quite significant since most of the sessions presenting high values for this combination of responses resulted in satisfactory VOM outcomes. However, since the number of samples is, in general, small, all responses vary in their performance depending on the grouping hypothesis, despite the low level of presence of the 3-value among the quantitative responses. This means that the uncertainty of the mediator when assigning a value of 3 to the answers tends to add noise to the overall data with respect to the evaluation.

The result tables show that CF and FF (and even LDA) vary significantly in their predictions depending on whether the nervousness features $\{f_{35}, f_{36}\}$ are considered or not. This indicates that the subjective evaluation of the mediator adds an important weight to the system for half of our classifiers. Moreover, the variability in performance presented by the remaining classifiers in relation to these two cases leads us to analyze the relevance of these features in each case. Thus, we performed a comparison to identify the most relevant features for each social response. In this way, we also analyzed the influence of the nervousness features $\{f_{35}, f_{36}\}$ when choosing the

Weighted feature selection predicting the Satisfaction



Weighted feature selection predicting the Agreement



Weighted feature selection predicting the Receptivity

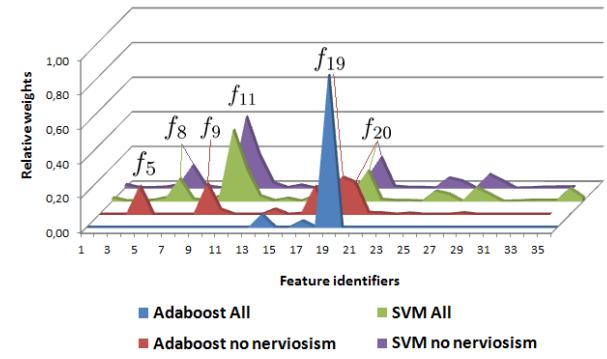


Fig. 9. Weighted feature selection when using AdaBoost and SVM for the grouping response cases presenting the highest accuracy when predicting receptivity, agreement, and satisfaction. Each line represents the relative feature weights assigned by the classifiers within the range $[0, 1]$, either employing all features or without the nervousness features f_{35} and f_{36} .

most relevant of the other features. We performed a weighted feature selection using [47] and [48] for AdaBoost and SVM, respectively. For each response (receptivity, agreement, and satisfaction), we selected those features only for the cases giving the highest degree of accuracy. See the different plots in Figure 9. In general, we observe that agitation features and the mediator's speaking turns are chosen as the most relevant features when predicting satisfaction. By contrast, the feature chosen as being most relevant for predicting agreement is the age of the mediator. In the case of receptivity, the fact of withholding the nervousness features results in the most significant changes in the feature selection with respect to the other responses. However, both hand agitation, gaze, and the

combination of the two are chosen as being the most relevant features when predicting receptivity. On average, the most relevant features for all the responses are those involving the combination of gaze and the agitation of the body regions. This means that these are the most discriminating behavioral indicators in the prediction of the degree of receptivity, agreement, and satisfaction in a conversation such as that maintained in a VOM process. This feature selection procedure has direct implications for the observational methodology of non-verbal communication, since it allows experts in the field of psychology and restorative justice to focus, in any given conversation, on the most discriminating behavioral indicators automatically selected through artificial intelligence.

Finally, we relate the overall training data to the different ground truth annotations using the epsilon-SVR regression strategy. In this case, when using the leave-one-out strategy, we obtain a prediction for each sample within the same range as the quantified annotations [1, 5]. In this setting, we also ran the experiment twice: first, we considered all features, and then left out the nervousness features. Both cases gave the same average distances when predicting satisfaction, agreement, and receptivity, with values of 0.59, 0.64, and 0.68, respectively. This mean deviation with respect to the ground truth labels was found to be significant and of great interest to the team of mediators.

V. CONCLUSION

We have proposed a multi-modal framework for the semi-automatic analysis of non-verbal communication in VOM sessions. We have demonstrated the usability of computer vision, signal processing, and machine learning strategies in conversational processes. Specifically, we computed a set of features using audio-RGB-depth data. Then, a heuristic procedure was presented within the multi-modal feature extraction to improve the continuity in the detection of positive regions of interest between consecutive frames. Finally, we have defined an automatic computation of behavioral indicators used as final features for learning and classification tasks. We have demonstrated the applicability of the system for use in the restorative justice field as a tool for mediators, obtaining recognition accuracies of 86% when predicting satisfaction and 79% when predicting both agreement and receptivity, and a high correlation in the regression analysis.

In future work, we plan to improve the dataset and responses, and to incorporate new features. In the case of the data, we hope to capture more samples so as to be able to perform more accurate predictions, providing continuous ground truth information by means of intra-mediator estimations. In the case of the predictions, new data should allow the continuous prediction of each degree of the behavioral indicators. Moreover, they will enable us to perform frame-based predictions, analyzing the evolution of each indicator throughout the VOM process, and to detect the exact instant (or "click" to use the term employed by the mediators) when a party accepts the possibility of reaching an agreement. Finally, we also hope to incorporate emotional state features, obtained from facial expressions [23] and audio data [49].

REFERENCES

- [1] E. Weitekamp, "Reparative justice," *European Journal on Criminal Policy and Research*, vol. 1, no. 1, pp. 70–93, 1993.
- [2] M. S. Umbreit, *The Handbook of Victim Offender Mediation: An Essential Guide to Practice and Research*. John Wiley & Sons, 2002.
- [3] A. S. Pentland, "Socially aware computation and communication," *Computer*, vol. 38, no. 3, pp. 33–40, 2005.
- [4] A. S. Pentland, *Honest Signals: How They Shape Our World*. Massachusetts: The MIT Press, 2008.
- [5] M. Knapp and J. Hall, *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 1997.
- [6] G. Mohammadi, S. Park, K. Sagae, A. Vinciarelli, and L.-P. Morency, "Who is persuasive?: The role of perceived personality and communication modality in social multimedia," in *ICMI*, 2013, pp. 19–26.
- [7] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. on Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.
- [8] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, and D. Gatica-Perez, "Body communicative cue extraction for conversational analysis," *FG*, 2013.
- [9] O. Aran and D. Gatica-Perez, "One of a kind: Inferring personality impressions in meetings," in *ICMI*, 2013, pp. 11–18.
- [10] A. Vinciarelli, H. Salamin, and M. Pantic, "Social signal processing: Understanding social actions through nonverbal behaviour analysis," in *CVPR*, vol. 3, 2009, pp. 42–49.
- [11] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.
- [12] D. B. Jayagopi and D. Gatica-Perez, "Mining group nonverbal conversational patterns using probabilistic topic models," *IEEE Trans. on Multimedia*, vol. 12, no. 8, pp. 790–802, 2010.
- [13] D. Sanchez-Cortes, O. Aran, D. Jayagopi, M. Schmid Mast, and D. Gatica-Perez, "Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 39–53, 2013.
- [14] S. Escalera, O. Pujol, P. Radeva, J. Vitrià, and M. T. Anguera, "Automatic detection of dominance and expected interest," *EURASIP Advances in Signal Processing, Research Article*, 2010.
- [15] T. Takahashi and F. Kishino, "Hand gesture coding based on experiments using a hand gesture interface device," *SIGCHI Bull.*, vol. 23, no. 2, pp. 67–74, 1991.
- [16] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *SMC-A*, vol. 37, no. 1, pp. 1–9, 2007.
- [17] V. Ponce, M. Gorga, X. Baró, and S. Escalera, "Human behavior analysis from video data using bag-of-gestures," in *IJCAI*, vol. 3, 2011, pp. 2836–2837.
- [18] J.-I. Biel and D. Gatica-Perez, "Vlogsense: Conversational behavior and social attention in youtube," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 7S, no. 1, pp. 33:1–33:21, 2011.
- [19] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press, 1992.
- [20] D. McNeill, *Gesture and Thought*. Chicago: University of Chicago Press, 2005.
- [21] A. Mehrabian, *Nonverbal communication*. Aldine-Atherton, 1972.
- [22] A. Hernández-Vela, M. Reyes, V. Ponce, and S. Escalera, "Grabcut-based human segmentation in video sequences," *Sensors*, vol. 12, no. 11, pp. 15 376–15 393, 2012.
- [23] O. Rudovic, M. Pantic, and I. Patras, "Coupled gaussian processes for pose-invariant facial expression recognition," *IEEE TPAMI*, vol. 35, no. 6, pp. 1357–1369, 2013.

- [24] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," *CVPR*, pp. 1297–1304, 2011.
- [25] A. Hernandez-Vela, M. Bautista, X. Perez-Sala, V. Ponce, X. Baro, O. Pujol, C. Angulo, and S. Escalera, "BoVDW: Bag-of-visual-and-depth-words for gesture recognition," *ICPR*, pp. 449–452, 2012.
- [26] O. Lopes, M. Reyes, S. Escalera, and J. Gonzalez, "Spherical blurred shape model for 3-d object and pose recognition: Quantitative analysis and hci applications in smart environments," *SMC-B: Cybernetics, IEEE Trans. on*, vol. PP, no. 99, 2014.
- [27] M. A. Bautista, A. Hernandez-Vela, V. Ponce, X. Perez-Sala, X. Baro, O. Pujol, C. Angulo, and S. Escalera, "Probability-based dynamic time warping for gesture recognition on rgbd data," *ICPR*, 2012.
- [28] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE TPAMI*, vol. 31, no. 7, pp. 1264–1277, 2009.
- [29] A. Stefan, V. Athitsos, J. Alon, and S. Sclaroff, "Translation and scale-invariant gesture recognition in complex scenes," in *PETRA*, 2008, pp. 7:1–7:8.
- [30] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," *Perceptual Comput. Sect.*, MIT, Cambridge, MA, USA, 1996.
- [31] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE TPAMI*, vol. 19, no. 7, pp. 780–785, 1997.
- [32] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The lium speech transcription system: a cmu sphinx iii-based system for french broadcast news," in *Interspeech*, 2005.
- [33] J. Ajmera, "Robust audio segmentation," Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, 2004.
- [34] X. Anguera and J. Pardo, "Robust speaker diarization for meetings: Icsi rt06s evaluation system," in *ICSLP*, 2006.
- [35] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
- [36] M. Fisher, "Interpreting sensor values." [Online]. Available: <http://graphics.stanford.edu/~mdfisher/Kinect.html>
- [37] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," *CVPR*, 2012.
- [38] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011.
- [39] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.
- [40] J. Dovidio and S. Ellyson, "Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening," in *SPQ*, 1982, pp. 106–113.
- [41] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE ICRA*, 2011.
- [42] S. Escalera, X. Baró, J. Vitrià, P. Radeva, and B. Raducanu, "Social network extraction and analysis based on multimodal dyadic interaction," *Sensors*, vol. 12, no. 2, pp. 1702–1719, 2012.
- [43] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, 1996, pp. 148–156.
- [44] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. IST*, vol. 2, pp. 27:1–27:27, 2011.
- [45] D. F. Specht, "Probabilistic neural networks for classification, mapping, or associative memory," in *IEEE IJCNN*, 1988, pp. 525–532 vol.1.
- [46] M. Gopikrishnan and T. Santhanam, "Effect of different neural networks on the accuracy in iris patterns recognition," in *IJRJC*, 2011, pp. 22–28 vol.7.
- [47] S. Escalera, O. Pujol, and P. Radeva, "Error-correcting ouput codes library," *JMLR*, vol. 11, pp. 661–664, 2010.
- [48] Y. wei Chen, "Combining svms with various feature selection strategies," in *Taiwan University*. Springer-Verlag, 2005.
- [49] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.