Pattern Recognition Letters 67 (2015) 19–27

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Non-verbal communication analysis in Victim–Offender Mediations☆

Víctor Ponce-López [a,b,c,*], Sergio Escalera [b,c], Marc Pérez [b], Oriol Janés [b], Xavier Baró [a,b,c]

[a] Internet Interdisciplinary Institute, Open University of Catalonia, Roc Boronat, 117, Barcelona 08018, Spain
[b] Department of Applied Mathematics and Analysis, University of Barcelona, Gran Via 585, Barcelona 08007, Spain
[c] Computer Vision Center, Autonomous University of Barcelona, Building O, 08193 Bellaterra, Barcelona, Spain

ARTICLE INFO

Article history:
Available online 2 September 2015

Keywords:
Victim–Offender Mediation
Multi-modal human behavior analysis
Face and gesture recognition
Social signal processing
Computer vision
Machine learning

ABSTRACT

We present a non-invasive ambient intelligence framework for the semi-automatic analysis of non-verbal communication applied to the restorative justice field. 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 Victim–Offender Mediation sessions in Catalonia. 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.

1. 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* [43]. 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 then taken as to whether the victim and the offender might engage in a joint encounter. Fig. 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 by the events that have occurred and the associated 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 [41]. 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, the 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 [28,29].

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. Indeed, no standard guidelines exist for establishing the best course of actions or identifying the psychological mechanisms for achieving the desired mediation goals. There exist, however, a set of body communicative cues that are present in the conversation and affect the way of how participants perceive each other. This non-verbal communication has been of high interest to intensively analyze the human interaction in social psychology and cognitive sciences [20].

In this context, multi-modal intelligent systems can be used to analyze this information by means of extracting features separately for

☆ This paper has been recommended for acceptance by Lledó Museros.

* Corresponding author. Tel.: +34934505254.
E-mail addresses: vponcel@uoc.edu, v88ponce@gmail.com (V. Ponce-López), sergio@maia.ub.es (S. Escalera), marcperez1993@gmail.com (M. Pérez), orioljanes@gmail.com (O. Janés), xbaro@uoc.edu (X. Baró).

http://dx.doi.org/10.1016/j.patrec.2015.07.040
0167-8655/

Fig. 1. 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

the different data sources, such as those captured from low-cost sensor devices. They can then be combined so as to define and recognize communicative indicators. In this paper, we present the first state of the art pattern recognition method to extract multi-modal features and recognize social signals in VOM processes.

The rest of the paper is organized as follows. Section 2 describes the related work. Next, Section 3 presents the material acquired and used in this study. In Section 4, we describe the system modules. Section 5 outlines the proposal setup and the experimental results. Finally, Section 6 concludes the paper.

2. Related work

The Restorative Justice approach focuses on the personal needs of victims. Achieving success in the VOM sessions depends largely on how the participants communicate with each other. [41] handbook 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 are of a highly sensitive nature, participants manifest emotional states when interacting with the others that can be physically observed through their non-verbal communication [20].

Recently, a number of studies have proposed ways in which personality traits can be inferred from multimedia data [27] 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 [36] 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 [3,22] combined several methodologies to analyze non-verbal behavior automatically by extracting communicative cues from both simulated and real scenarios. Additionally, information obtained from speech is commonly used [19,42], as is other information obtained from ambient and wearable sensors [35]. In [13], 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 [28,29].

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. However, because of the need to avoid wearing intrusive egocentric devices, some ambient sensors that provide multi-modal data might be considered. In [22], 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 [30], which provides trajectory analyses from body movements and gestures. Furthermore, audio information has been analyzed in [6], with the objective of modeling descriptors for speech recognition. Beyond these works, in this paper we propose another mid level of abstraction to obtain behavioral indicators based on communicative cues, which are able to better explain those features that are directly extracted from multi-modal data. Moreover, these behavioral features will be combined to describe additional behavioral indicators and analyze their influence in VOM scenarios.

The analysis of the participants from a computer vision point of view use to be defined by region of interest detection, description, and tracking, usually involving the face or hands. These regions provide discriminative behavioral information, or adaptors, which are movements, such as head scratching, indicative of attitude, anxiety level and self-confidence [24]; 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 [25]. However, as explained in [22,26], 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 [33,42].

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 3*D* partial information to be obtained from the scene. In [37], the authors proposed a system for real-time human pose recognition including depth information for each image pixel. This new source of information has been recently exploited for creating new human pose descriptors by combining different state of the art RGB-depth features [18], as well as they are used in a large amount of Human Computer Interaction (HCI) applications [21].

Once body features are computed, behavioral indicators can be analyzed by studying their trajectories using pattern recognition approaches. Some of the methods in this context are based on dynamic programming techniques such as Dynamic Time Warping (DTW) [18]

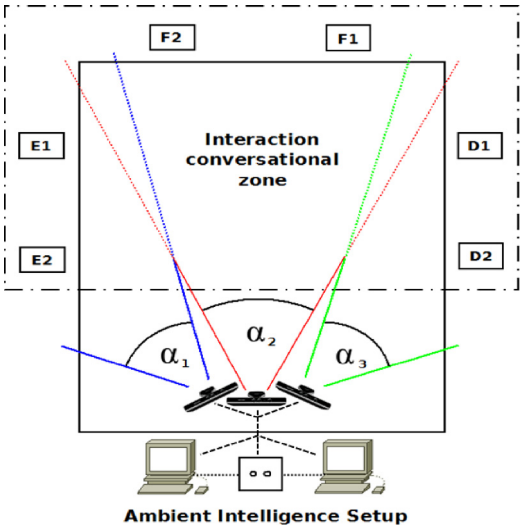

Fig. 2. Acquisition architecture. *E*1, *E*2, *F*1, *F*2, *D*1, *D*2 are the participants codified by their respective positions (E: left, F: front, D: right); the angles of view for the different cameras are the same, and hence $\alpha_1 = \alpha_2 = \alpha_3$.

or involve statistical approaches, such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [39,40,45].

Once data from the environment have been acquired and processed to define a set of behavioral features, they serve as the basis for modeling a set of communication indicators. For instance, in [44], the authors outline a system for real-time tracking of the human body with the objective of interpreting human behavior.

In this paper, we propose the first state of the art non-invasive ambient intelligence framework for the semi-automatic analysis of non-verbal communication in VOM processes. 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.

3. Data collection

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 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 recorded.[1] Fig. 2 shows the ambient intelligence setup with all the elements involved and their distribution.

Recordings were made in various towns and cities of Catalonia. Most of them 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 were recorded, with a duration from 20 min to 2 h depending on the session, and an overall average of 35 min among all sessions. For each session, 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. 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 session[2] provided audio-RGB-depth information. These modalities were registered using the camera parameters, 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 m depending on 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 there were 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 2 [3,13,19,22,27–29,35,36,41,42], 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).

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 1 shows a numerical summary of the data acquired.

4. Proposed method

The proposed framework consists of three main sequential modules illustrated in Fig. 3. The first module includes the multi-modal feature extraction from audio-RGB-depth data. The steps for obtaining multi-modal features from different sources of information are the speaker diarization, user segmentation, and region detection.

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

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

Table 1
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^o participants	37
Total n^o frames	1,436,400
Average n^o minutes/session	35

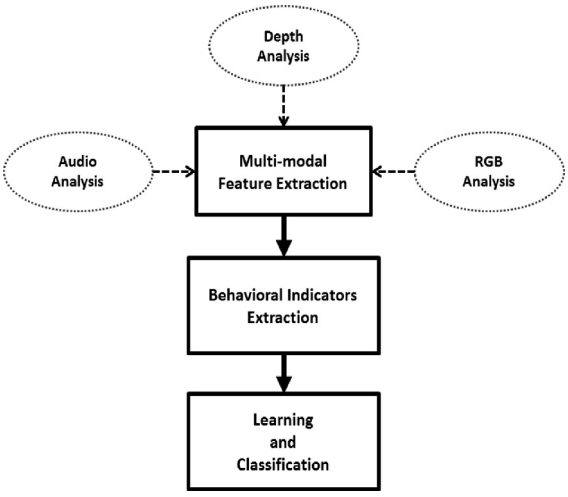

Fig. 3. Modules of the proposed system.

Once the multi-modal features have been extracted, they are used to define the behavioral indicators to be learnt and classified.

The remainder of this section describes the different parts of the multi-modal feature extraction block of Fig. 3, followed by the behavioral indicators, and finally the learning and classification of receptivity, agreement, and satisfaction labels.

4.1. Audio analysis: speaker diarization

In order to obtain the audio features, we use a diarization scheme based on the approach presented in [9]. These features correspond to state of the art methods for audio descriptions, which have been successfully applied in several audio analysis applications [1,2,32]. 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:

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

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 $\tilde{x}(n)$,

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

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. 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.

4.2. User detection

Both RGB and depth data are used for the postural and behavioral analyses of the parties. In this sense, the first step involves performing a limb-segmentation of the body based on the Random Forest method of [37]. Fig. 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 [15], 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 [15], 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:

$$x = (\dot{x} - \delta_x)\frac{d(\dot{x},\dot{y})}{\kappa_x}, y = (\dot{y} - \delta_y)\frac{d(\dot{x},\dot{y})}{\kappa_y}, z = d(\dot{x},\dot{y}), \tag{3}$$

where (x, y, z) will be the real world coordinates, and $\delta_x, \delta_y, \kappa_x, \kappa_y$, 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.

4.3. 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.

4.3.1. 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 [46] which uses a set of face models. The face model

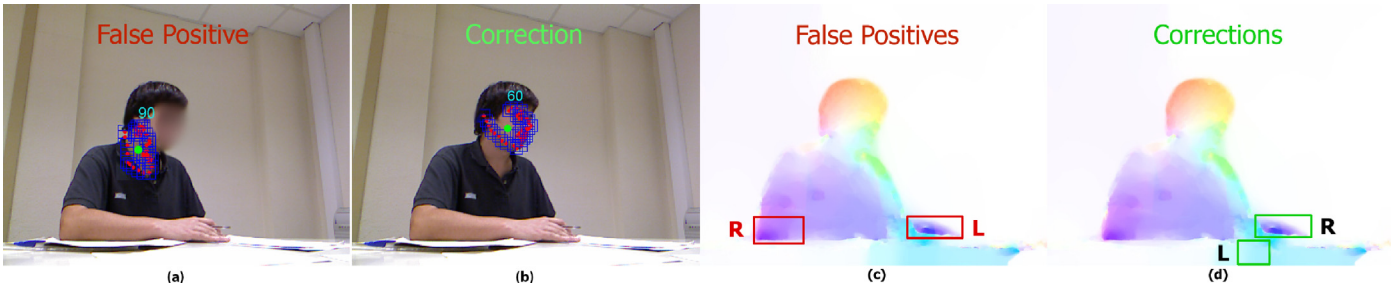

Fig. 4. Examples of how the semi-automatic heuristic procedure of [31] works on two pairs of frames of a session. The correction of false positives is shown, improving the continuity of the detection of positive regions of interest between consecutive frames. Image (a) shows a false positive detection 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, recovered from the previous frames.

is based on a mixture of trees with a shared pool of parts, where every facial landmark is modeled 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. On the other hand, 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 [14]. In Fig. 1(a) we can visualize the set of computed features plotted on the detected face.

While face detection takes place for each tested image, we use the semi-automatic heuristic procedure of [31] so as 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. Fig. 4(a) and (b) shows an example of correcting a false positive detection.

4.3.2. Hand analysis

Given that the skeletal model computed from the person segmentation image [37] 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. The bounding boxes of Fig. 4(c) show an example of detections (left is incorrect) using this procedure.

To improve the detection, we use the same heuristic procedure as that applied to the face analysis step for choosing, in this case, the two best hand candidates. Image (d) of Fig. 4 shows 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 Eq. (3).

4.3.3. Upper body analysis

As presented above, the probability of each pixel of an image belonging to a labeled body part is computed using depth features. This information is used for the subsequent calculation of optical flow on RGB images where the upper body region appears. Therefore, each pixel $\dot{p}$ 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, which is later used to define behavioral indicators.

4.4. Behavioral indicators

Once the multi-modal features have been extracted, we use them to build a set of behavioral indicators that reveal communicative cues. This set of behavioral indicators defines the final feature vector for each party within the VOM process. This information is of great interest in detecting the response of subjects to certain feelings or emotional states during the conversation [20]. 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.

4.4.1. Target gaze codification

The head pose and the face is obtained by applying the methodology explained in Section 4.3.1. In a given session, we compute the correlations between the head pose angles belonging to each participant and the positions taken by the remainder participants in that session. Hence, we identify the visual focus of attention among the different participants in the conversation [3,5,23]. 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 Fig. 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

Table 2
Summary of behavioral indicators defining each feature vector. The last two features derive from the mediator surveys.

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}	Hands together of this party
f_{23}	Hands of this party touching the face
f_{24}	Hands of this party are 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

depending on each session setup. Then, the target gazes are automatically identified for all the frames of the session.

Fig. 1(a) shows an example of crossed gazes between the mediator and a party in a real VOM session. Finally, we compute the time percentages of target gazes for each party. Therefore, for any given party, there is a total of 6 indicators for representing the target gazes ($\{f_{15}, f_{16}\}$ and $\{f_{18} - f_{21}\}$ from Table 2).

4.4.2. Agitation estimation

As explained in Section 4.3.2, 3*D* 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 $F = \{\iota_1, \iota_2, \iota_3, \ldots, \iota_\lambda\}$ 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:

$$A_h = \frac{1}{\lambda} \sum_{\iota=1}^{\lambda} \Delta_h^\iota, \tag{4}$$

where $\Delta_h^\iota = \Delta_p^\iota + \Delta_q^\iota$ are the displacements among 3*D* 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 4.3.3 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:

$$A_b = \frac{1}{\lambda} \sum_{\iota=1}^{\lambda} \bar{\sigma}_\iota, \tag{5}$$

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 ($\{f_{14} - f_{21}\}$ from Table 2), either alone or in combination with other indicators. The idea of combining these indicators with other behavioral features is inspired by [10,13]. 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, as in [31].

4.4.3. Posture identification

From the 3*D* body position, we detect the body posture as one 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, 3*D* 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 3*D* 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 a threshold. 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 belonging to the face center of mass obtained in Section 4.3.1. Then, we consider the frames where the distance values are below that of the threshold. 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 [34] to obtain planar objects within images. Then, we compare the 3*D* 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. Fig. 1(b) and (c) illustrate an example of this procedure, showing respectively the input depth image and its segmentation.

4.4.4. Speech turns/interruptions detection

The speaker diarization process of Section 4.1 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 [11]. Using the time between turns, we compute the percentage of turns in which a participant interrupts another one.

4.5. Classification

The total number of behavioral indicators is 36 (see [31]), 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. Thus, if a session 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. Each party of a video session is a sample for the classification task, and the total number of used samples is 28.

As explained in Section 3, 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 [16], Support Vector Machines (SVM) using a Radial Basis Function (RBF) [7], Linear Discriminant Analysis (LDA), and three kinds of Artificial Neural Networks (ANN), in particular Probabilistic Neural Networks (PNN) [38], and Cascade-Forward (CF) and Feed-Forward neural networks (FF) [17]. In addition to the binary classification analysis we also conduct a regression study using epsilon-SVR (Support Vector Regression) [7] in order to predict continuous quantifications of the three labels.

5. Experiments

5.1. Setting and validation measurements

The measurements for the features referring to the gaze, interaction of hands, and the position of hands respect to the table, are time percentages. The features referring to agitation, combination of agitation and gazes, and interaction of hands with the face, contain averaged values of optical flow or distances, all of them taking into account the processed frames of a session. The features referring to the speech are turn taking percentages, where a turn means that the speaker changes. Finally, the remaining features, including nervousness features, are codified either into binary values or discrete values within a certain range, having 5 as the maximum range length.

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.

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 grouping case, as we show in the result Tables 3–6:

- First grouping case: Degrees of quantification {1, 2, 3} versus {4, 5}.
- Second grouping case: Degrees of quantification {1, 2} versus {3, 4, 5}.

Table 3
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 4
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%

Table 5
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 6
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%

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.

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.

5.2. Results

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 3 and 4 show the results obtained when employing the different techniques

and using the complete set of behavioral indicators of Table 2. Note that as the features of nervousness are subjective indicators that are not automatically computed, we repeated the experiments without these two features. These results are shown in Tables 5 and 6, where the prediction is also analyzed under the grouping hypotheses. The most accurate results among the four tables for the three responses are shown in bold, showing both which classifier and which grouping case give the best performance for each feature description. 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 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 when choosing the most relevant of the other features. We performed a weighted feature selection using [12] and [8] 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 Fig. 5). 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

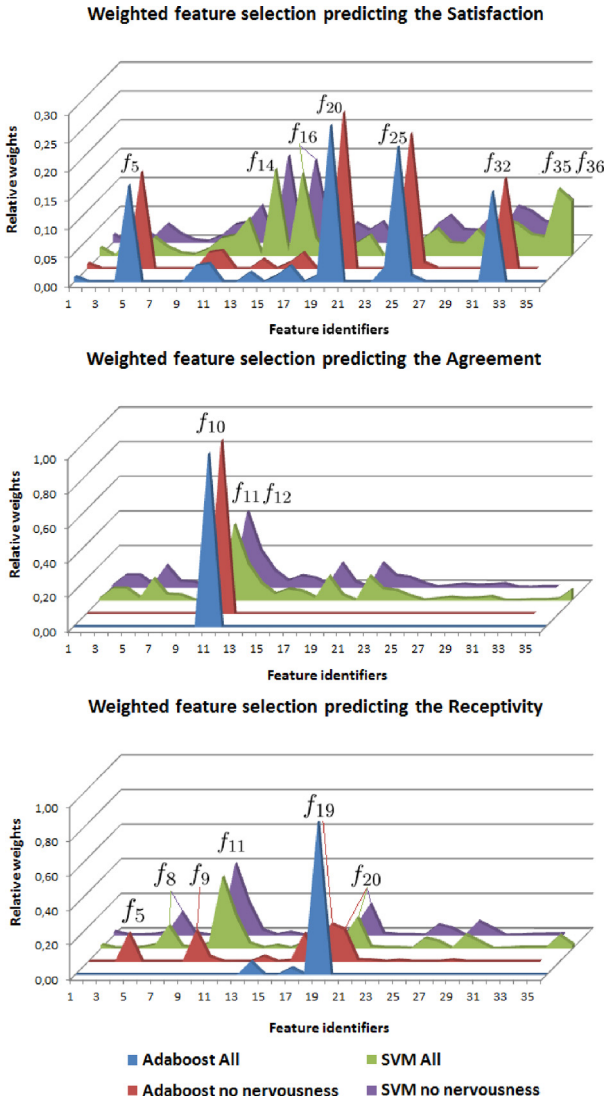

Fig. 5. Weighted feature selection when using AdaBoost and SVM for the grouping response cases presenting the highest accuracy when predicting receptivity, agreement, and satisfaction. Each line represents the relative feature weights assigned by the classifiers within the range [0, 1], either employing all features or without the nervousness features f_{35} and f_{36}.

[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 similar 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 accurate and of interest to the team of mediators.

6. Conclusion

We proposed a multi-modal framework for the semi-automatic analysis of non-verbal communication in VOM sessions. We showed the usability of computer vision, signal processing, and machine learning strategies in conversational processes. Specifically, we computed a set of multi-modal features from multimodal data. Then, we defined an automatic computation of behavioral indicators used as final features for learning and classification tasks. We demonstrated the applicability of the system to be used in the restorative justice field as a tool for mediators, obtaining recognition accuracies of 86% when predicting satisfaction, 79% when predicting both agreement and receptivity, and a high correlation in the regression analysis.

As 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, it will enable us to perform frame-based predictions, analyzing the evolution of each indicator throughout the VOM process, and to detect the exact instant when a party accepts the possibility of reaching an agreement. Finally, we plan to incorporate emotional state features obtained from facial expressions [33] and audio data [4].

Acknowledgments

This project is supported by the projects TIN2012-38187-C03-02, TIN2013-43478-P, and the grant 2013FI-B01037 from the Catalan government.

References

[1] J. Ajmera, I.A. McCowan, H. Bourlard, Robust Audio Segmentation, École Polytechnique Férale de Lausanne, Switzerland, 2004 Ph.D. thesis.
[2] X. Anguera, J. Pardo, Robust speaker diarization for meetings: ICSI RT06S evaluation system, in: Proceedings of the International Conference on Spoken Language Processing, ICSLP, 2006.
[3] O. Aran, D. Gatica-Perez, One of a kind: inferring personality impressions in meetings, in: Proceedings of the ICMI, 2013, pp. 11–18, doi:10.1145/2522848.2522859.
[4] M.E. Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recognit. 44 (3) (2011) 572–587.
[5] S. Ba, J.-M. Odobez, Recognizing visual focus of attention from head pose in natural meetings, Syst., Man, Cybernet.-B 39 (1) (2009) 16–33.
[6] J.-I. Biel, D. Gatica-Perez, VlogSense: conversational behavior and social attention in Youtube, ACM TOMCCAP 7S (1) (2011) 33:1–33:21.
[7] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. IST 2 (2011) 27:1–27:27.
[8] Y. wei Chen, Combining SVMs with Various Feature Selection Strategies, Springer-Verlag, Taiwan University, 2005.
[9] P. Deléglise, Y. Estève, S. Meignier, T. Merlin, The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news, in: Proceedings of the Interspeech, 2005.
[10] J. Dovidio, S. Ellyson, Decoding visual dominance: attributions of power based on relative percentages of looking while speaking and looking while listening, in: Proceedings of the SPQ, 1982, pp. 106–113.
[11] S. Escalera, X. Baró, J. Vitrià, P. Radeva, B. Raducanu, Social network extraction and analysis based on multimodal dyadic interaction, Sensors 12 (2) (2012) 1702–1719, doi:10.3390/s120201702.
[12] S. Escalera, O. Pujol, P. Radeva, Error-correcting output codes library, J.Mach. Learn. Res. 11 (2010) 661–664.
[13] S. Escalera, O. Pujol, P. Radeva, J. Vitrià, M.T. Anguera, Automatic detection of dominance and expected interest, EURASIP Adv. Signal Process., Research Article (2010).
[14] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[15] M. Fisher, Interpreting sensor values, http://graphics.stanford.edu/mdfisher/Kinect.html.
[16] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the ICML, 1996, pp. 148–156.
[17] M. Gopikrishnan, T. Santhanam, Effect of different neural networks on the accuracy in iris patterns recognition, in: Proceedings of the IJRIC, vol. 7, 2011, pp. 22–28.
[18] A. Hernández-Vela, M.-A. Bautista, X. Perez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol, C. Angulo, Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in RGB-D, Pattern Recognition Letters 50 (2014) 112–121.
[19] D.B. Jayagopi, D. Gatica-Perez, Mining group nonverbal conversational patterns using probabilistic topic models, IEEE Trans. Multimed. 12 (8) (2010) 790–802, doi:10.1109/TMM.2010.2065218.
[20] M. Knapp, J. Hall, Nonverbal Communication in Human Interaction, Harcourt Brace College Publishers, 1997.
[21] O. Lopes, M. Reyes, S. Escalera, J. Gonzalez, Spherical blurred shape model for 3-d object and pose recognition: quantitative analysis and HCI applications in smart environments, Syst., Man, Cybernet.-B PP (99) (2014).
[22] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, D. Gatica-Perez, Body communicative cue extraction for conversational analysis, Automatic Face and Gesture Recognition (FG) (2013).
[23] M. Marin-Jimenez, A. Zisserman, M. Eichner, V. Ferrari, Detecting people looking at each other in videos, Int. J. Comput. Vis. 106 (3) (2014) 282–296.
[24] D. McNeill, Hand and Mind: What Gestures Reveal about Thought, University of Chicago Press, 1992.
[25] D. McNeill, Gesture and Thought, University of Chicago Press, 2005.
[26] A. Mehrabian, Nonverbal communication, Aldine-Atherton, 1972.
[27] G. Mohammadi, S. Park, K. Sagae, A. Vinciarelli, L.-P. Morency, Who is persuasive?: the role of perceived personality and communication modality in social multimedia, in: Proceedings of the ICMI, 2013, pp. 19–26.
[28] A.S. Pentland, Socially aware computation and communication, Computer 38 (3) (2005) 33–40, doi:10.1109/MC.2005.104.
[29] A.S. Pentland, Honest Signals: How They Shape Our World, The MIT Press, Massachusetts, 2008.
[30] V. Ponce, M. Gorga, X. Baró, S. Escalera, Human behavior analysis from video data using bag-of-gestures, in: Proceedings of the IJCAI, vol. 3, 2011, pp. 2836–2837.
[31] V. Ponce-López, S. Escalera, X. Baró, Multi-modal social signal analysis for predicting agreement in conversation settings, in: Proceedings of the ICMI, ACM, New York, NY, USA, 2013, pp. 495–502.
[32] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
[33] O. Rudovic, M. Pantic, I. Patras, Coupled Gaussian processes for pose-invariant facial expression recognition, IEEE TPAMI 35 (6) (2013) 1357–1369.
[34] R.B. Rusu, S. Cousins, 3D is here: point cloud library (PCL), in: Proceedings of the IEEE ICRA, 2011.
[35] D. Sanchez-Cortes, O. Aran, D. Jayagopi, M. Schmid Mast, D. Gatica-Perez, Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition, J. Multimodal User Interfaces 7 (1-2) (2013) 39–53.
[36] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, D. Gatica-Perez, A nonverbal behavior approach to identify emergent leaders in small groups, IEEE Trans. Multimed. 14 (3) (2012) 816–832.
[37] J. Shotton, A.W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, Comput. Vis. Pattern Recognit. (2011) 1297–1304.
[38] D.F. Specht, Probabilistic neural networks for classification, mapping, or associative memory, in: Proceedings of the IEEE IJCNN, vol.1, 1988, pp. 525–532.
[39] T. Starner, A. Pentland, Real-time american sign language recognition from video using hidden markov models, IEEE Trans. Pattern Anal. Mach. Intell (1998) 1371–1375.
[40] A. Stefan, V. Athitsos, J. Alon, S. Sclaroff, Translation and scale-invariant gesture recognition in complex scenes, in: Proceedings of the PETRA, 2008, pp. 7:1–7:8.
[41] M.S. Umbreit, The Handbook of Victim Offender Mediation: An Essential Guide to Practice and Research, John Wiley & Sons, 2002.
[42] A. Vinciarelli, H. Salamin, M. Pantic, Social signal processing: understanding social interactions through nonverbal behaviour analysis, in: Proceedings of the CVPR, vol. 3, 2009, pp. 42–49.
[43] E. Weitekamp, Reparative justice, Eur. J. Crim. Policy Res. 1 (1) (1993) 70–93, doi:10.1007/BF02249525.
[44] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: real-time tracking of the human body, IEEE TPAMI 19 (7) (1997) 780–785.
[45] H.-D. Yang, S. Sclaroff, S.-W. Lee, Sign Language Spotting with a threshold model based on conditional random fields, IEEE TPAMI 31 (7) (2009) 1264–1277.
[46] X. Zhu, D. Ramanan, Face detection, pose estimation and landmark localization in the wild, Comput. Vis. Pattern Recognit. (2012).