Multi-modal Social Signal Analysis for Predicting Agreement in Conversation Settings

Víctor Ponce-López *
IN3, Open University of Catalonia, Roc Boronat, 117, 08018 Barcelona, Spain.
Dept. MAiA, University of Barcelona, Gran Via, 585, 08007 Barcelona, Spain.
Computer Vision Center, UAB, 08193 Barcelona, Spain.
vponcel@uoc.edu

Sergio Escalera
Dept. MAiA, University of Barcelona, Gran Via, 585, 08007 Barcelona, Spain.
Computer Vision Center, UAB, 08193 Barcelona, Spain.
sergio@maia.ub.es

Xavier Baró
EIMT, Open University of Catalonia, Rbla. Poblenou, 156, 08018 Barcelona, Spain.
Computer Vision Center, UAB, 08193 Barcelona, Spain.
xbaro@uoc.edu

ABSTRACT

In this paper we present a non-invasive ambient intelligence framework for the analysis of non-verbal communication applied to conversational settings. In particular, we apply feature extraction techniques to multi-modal audio-RGB-depth data. We compute a set of behavioral indicators that define communicative cues coming from the fields of psychology and observational methodology. We test our methodology over data captured in victim-offender mediation scenarios. Using different state-of-the-art classification approaches, our system achieve upon 75% of recognition predicting agreement among the parts involved in the conversations, using as ground truth the experts opinions.

1. INTRODUCTION

In a conversation setting with different people involved we find a set of communicative cues. These cues appear implicitly during the conversational process and provide rich information that may be missed by the observer. Often, the goal in a conversation is to achieve the agreement among the parts involved by obtaining an equilibrium among their arguments and interests. Moreover, high levels of agreement among the parts are determinant indicators to detect the success of a conversation [1].

In the fields of psychology and observational methodology one can find biological, psycho-social, and environmental factors that help to better understand what and how they affect to participants mood. Therefore, if these techniques were studied and established as self-knowledge tools for both communication and intervention, experts could obtain feedback from the conversations to analyze the process, and guide it toward it success.

In this context, multimedia data analysis from different sources can be used to achieve those objectives. This information can be analyzed on different modules in order to extract features separately, and later combine them to define communicative indicators.

1.1 Related work

Recently, many social signal analyses from multimedia data have been performed on small group conversations [2], where the goal is to detect and emphasize the importance of social signals in group interactions. In particular, these works present several approaches for automatically analyze non-verbal behavior by extracting body communicative cues in both simulated and real scenarios. Most of these social signal processing frameworks are used to detect a set of visual indicators (including body and face analysis), or including information obtained from the speech or other ambient and wearable sensors [3, 4].

Many of previous works proved that the agreement during the communication is an indicator highly dependent on social signals. Therefore, one is able to perform an exhaustive analysis to detect which are the roles of each participant in terms of influence, dominance, or submission. For instance, in [5], both the interest of observers and the dominant people are predicted using only behavioral motion information when looking at face-to-face (also called *vis-à-vis* or dyadic) interactions. Furthermore, there exist many interdisciplinary works in the state-of-the-art covering related fields under the social computing point of view, some of them summarized in [1].

In most of these social signal processing frameworks, one can observe that both ambient intelligence and egocentric computing methods are defined. In fact, this definition refers to the data acquisition procedure which mainly depends on how data captured from different devices is handled, as well as what devices are used for capturing these data. Ambient intelligence regards to electronic environments that are sensitive and responsive to the presence of people, whereas egocentric computing refers to the use of wearable devices. Therefore, the use of these devices in this procedure depends

*V. Ponce-López acknowledges a FI-DGR fellowship from the Catalan Government (2013FI-B01037). The work is also supported by the TIN2012-38187-C03-02 project.

ICMI'13, December 9–13, 2013, Sydney, Australia

http://dx.doi.org/10.1145/2522848.2532594.

on the environmental conditions of the application (scenario, people involved, the electronic distribution and organization of devices, or the intrusiveness of wearable devices). Often, existing techniques used for data acquisition make use of interface devices [6], or special wears like gloves [7] to increase recognition accuracy. However, while these techniques give striking results in simulated environments, their use becomes not feasible in real-case scenarios due to their invasiveness and the uncontrolled nature of the application context.

Because of the need to avoid wearing intrusive egocentric devices, some other ambient sensors that provide multimodal data can be considered. In [2], a custom developed system is applied in a real-case scenario for job interviews. The data acquisition procedure is performed using different types of cameras by setting them in different positions and ranges for capturing visual and depth information. Similarly, scenes with non-invasive systems have been proposed in other works like [8], which provides trajectory analyses from body movements and gestures. Furthermore, audio information has been analyzed in [9], with the objective of modeling descriptors for speech recognition. This can be useful information to measure the levels of activity from speech cues like detection of speech/non-speech, interruptions, pauses, or segments obtained from a speech diarization process.

In most of previous conversational contexts people usually appear either sitting or with some body parts occluded. Therefore, from a computer vision point of view one may be interested in focusing only on the upper body regions. Afterwards, feature extraction techniques can be performed using computer vision techniques, where sources of most significative information come from interest regions such as face and hands. These regions provide discriminative behavioral information called adaptors, which are movements like head scratching, attitude, anxiety level and self-confidence; and beat gestures, which are flicks of hands used to emphasize important parts of the speech with respect to the larger discourse [10]. However, as it is explained in [2, 11], body posture is also found to be an important indicator of the emotional state of a person. Additionally, another potential source of information is provided by facial expressions [4].

In most of the presented scenarios, classic computer vision techniques are applied on RGB data. It implies that either multiple cameras are installed and synchronized to obtain $3D$ data and other discriminant information, or some relevant behavioral and postural information is not captured. In this sense, recent works included compact multi-modal devices which allow to obtain $3D$ partial information from the scene. In [12], 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 Kinect$^{\text{TM}}$ 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 that provides visual $2.5D$ features has been recently applied by several authors. As an example, in [13] a new human pose descriptor is presented by combining different state-of-the-art RGB-depth features.

Many recent methods for detecting the body posture perform spatial optimization of their parts. In [12] this is performed using GrabCuts, which are based on multiple representations of Gaussian Mixture Models (GMM) and graph cuts optimization. Once body postures and their parts are represented, behavioral indicators are usually analyzed by studying their trajectories using both machine learning and pattern recognition approaches. Some methods in this context are based on dynamic programming techniques such as Dynamic Time Warping (DTW) [14] or statistical approaches such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) [15, 16].

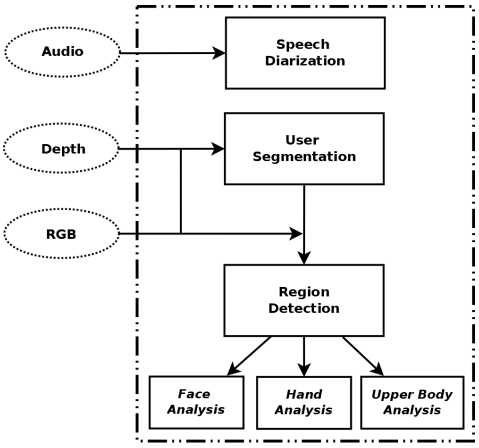

Figure 1: Multi-modal feature extraction module.

Once data from the environment are acquired and processed defining a set of behavioral features, these are the basis for modelling a set of communication indicators. In the context of conversations we are specially interested in behavioral traits belonging to social signals presented within both the communication and interactions among participants. In this sense, levels of activity, stress, and involvement can be analyzed not only from body movements, but also from the speech, facial expressions, or look. Therefore, we can then be able to perform an exhaustive analysis for modelling those significative social signals to estimate the levels of dominance or agreement in conversations.

1.2 Contributions

In this paper, we propose a non-invasive ambient intelligence framework for the analysis of real conversation settings. First, we extract a set of multi-modal audio-RGB-depth features from the analysis of faces, gestures, and speech, and an semi-automatic heuristic procedure is presented to correct and improve the continuity of positive region of interest detections. Then, we compute a set of behavioral communication indicators from the multi-modal features, which are used to measure the degree of agreement using state-of-the-art machine learning approaches and the ground truth defined in the data set. In particular, we use real data from victim-offender mediation processes, using as ground truth the level of agreement achieved based on the opinion of the experts. As a result, we found that our technology achieves good correlation between those relevant features regarding to behavioral indicators and the information provided by the experts.

2. PROPOSED METHOD

The proposed method for the non-verbal communicative analysis is described in this section. The framework consists of three main sequential modules. The first includes the multi-modal feature extraction from audio-RGB-depth

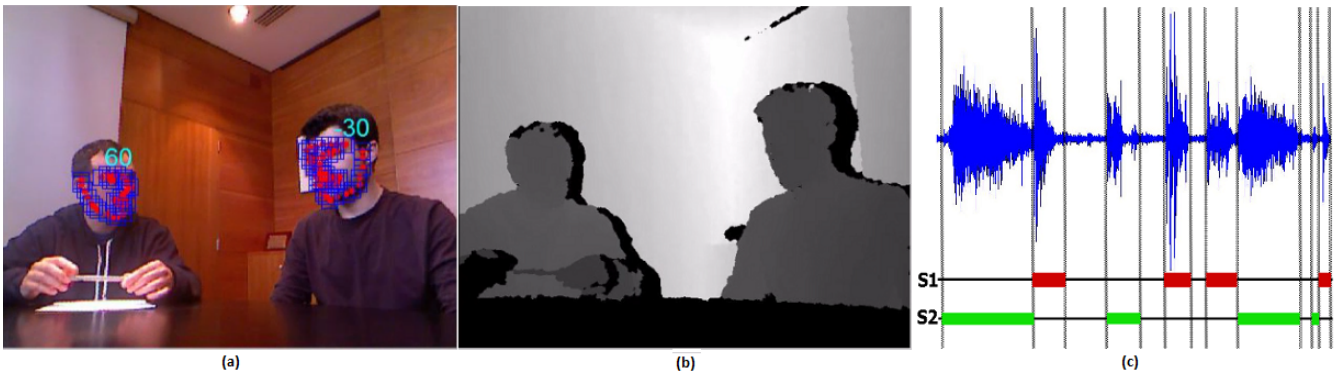

Figure 2: Example of the conversation scenario. Image (a) shows the RGB image with both face detection and head pose estimation. Image (b) shows the corresponding depth map, where the gray hues indicate the distance respect to the camera. Image (c) illustrates the speech diarization process for each subject S1 and S2, where clusters that belong to each one are obtained from the input signal estimating the speech time of each segment, as well as the speech pauses/interruptions.

data, which is summarized in Figure 1. As shown in the scheme, the steps for obtaining multi-modal features from different sources of information are the speech diarization, user segmentation, and region detection. Once multi-modal features are extracted, they are used to define the behavioral indicators in the second module, which are used as inputs for the learning and classification step in the third module of the system.

2.1 Conversation Setup

Figure 2 represents an example of a conversation scenario. In the next sections we refer to a session as a conversational setting where different individuals appear. Individuals (or subjects) are both men and women, adults, and belong to different parts or roles: victim, offender, or mediator in our particular victim-offender scenarios. For each session and part, the mediator fills a survey with information regarding his/her impressions about the conversation such as the agreement reached. The agreement answers are later used as ground truth for the classification in order to analyze the achieved agreement recognition rates.

2.2 Audio Analysis

This section describes the feature extraction applied to audio data source. As audio cues are the primary source of information, they can provide useful evidence for speech production.

2.2.1 Speech Diarization

In order to obtain the audio features, we use a diarization scheme based on the approach presented in [17]. These features correspond to the state-of-the-art on audio descriptions, which have been successfully applied in several audio analysis applications [18, 19]. The process is described next:

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. Finally, the Discrete Cosine Transform (DCT) is used in order to obtain the first 12 MFCC coefficients. Those coefficients are complemented with the energy coefficient and the dynamic features *delta* and *delta-delta*, which correspond to the first and second time-derivatives of Cepstral coefficients.

Speaker segmentation: Once the audio data is properly codified by means of those features, next step is to identify 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 full covariance matrix over the extracted features. This process produces an over segmentation of the audio into homogeneous small blocks. Then, a hierarchical clustering is applied over 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 Bayesian Information Criterion (BIC) stopping criterion is met. As in the previous step, each cluster is modeled by means of a Gaussian distribution with full covariance matrix and centroid distance is used as link similarity. Finally, a Viterbi decoding is performed in order to adjust the segment boundaries. Clusters are modelled by mean of a one state HMM having as observation model GMM with diagonal covariance matrices. Since most of people appear only in a single session, we do not learn speaker models from cluster GMMs. Therefore, models extracted from a session are not used on the diarization process of other sessions. Figure 2 (c) illustrates the procedure and the resulting segments assigned to the two subjects (two parts involved in the conversation).

2.3 User Detection

Both RGB and depth data are used for postural and behavioral analysis of the person. In this sense, the first step is to perform a limb-segmentation process of the body based on the Random Forest method of [12].

Once interest regions are located, it is of special interest to get the real-world distance values for some computed features in order to make them comparable among different subjects. For this, we inspired on a similar procedure as explained in [20], which converts the $2D$ pixels onto $3D$ real-world coordinates using the depth values provided by the Kinect^{TM} device. Figure 2 (b) shows the depth image corresponding to the RGB image (a).

2.4 Region Detection

This section describes the different modules for the feature extraction applied to visual data source once user has been segmented. In particular, we perform an analysis of the face, hands, and upper body, as well as visual movements performed by these regions during conversations. Moreover, the estimation of head/body postures provides information about both where each person may be looking at, and the status of the people in terms of agitation.

2.4.1 Face Analysis

Face analysis is the first feature extraction step, where we are mainly interested in obtaining the head pose angle for each person. First, we base on the method of [21]. It contains a set of previously trained faces in order to create the 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.

While face detection takes place for each tested image, a semi-automatic heuristic procedure is proposed in order to both improve the continuity of positive detections on the person among consecutive frames, and correct possible miss-detections due to the inherent difficulties of the problem at hand. Figure 3 shows the flowchart of the procedure applied to each frame. Summarizing, it consists of a temporal filtering methodology of detected regions (faces) among one-by-one consecutive frames. It is based on three main constraints that allow to choose the detected regions at the current frame in comparison to the 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 positive detections $h^\wp$, false positives $FP^\wp$ and false negatives $FN^\wp$. Then, a confidence ζ is computed from $h^\wp$ and the sum of false detections $\varepsilon^\wp$ 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 to be 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 both assure the continuity of positive detections for that person (even though the method could not detect any region in the image), and decide whether a manual annotation is required to re-initialize the detection process in the (approximately) desired frequency rate. The RGB image (a) of Figure 2 shows the two meshes obtained after this procedure fitted the face regions and the estimation of their head poses (in terms of orientation degrees).

2.4.2 Hand Analysis

Given that the skeletal model computed from the person segmentation image [12] does not offer accurate fitting of the hand joints in our particular scenario, we designed a semi-automatic procedure for hand detection, considerably

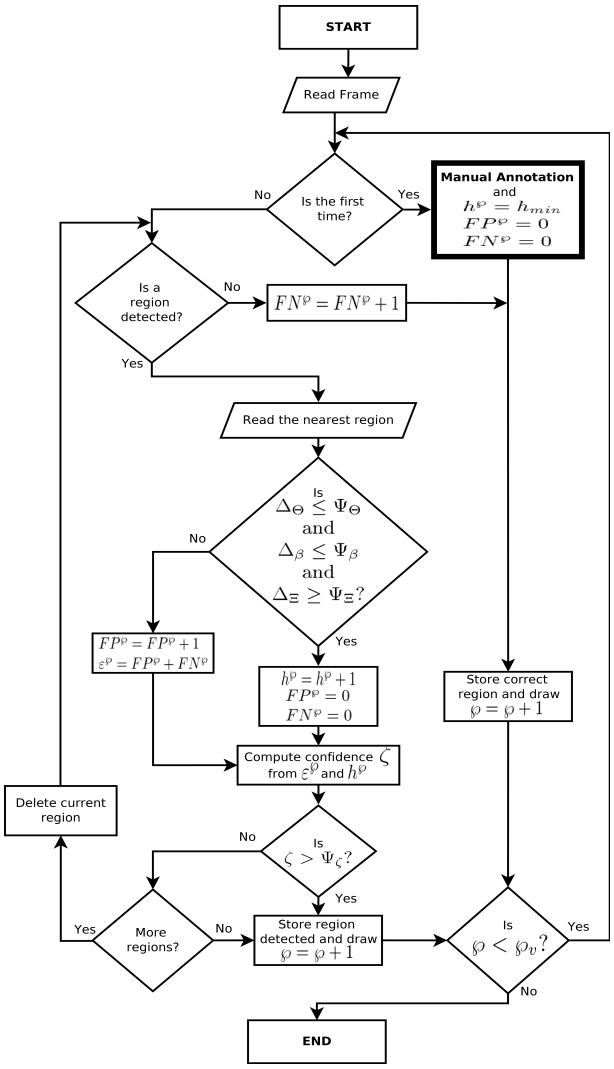

Figure 3: Flowchart of the heuristic procedure applied to each frame. The total number of people that appear in the current video is denoted by $\wp_v$. Constraints of the main condition at the center of the flowchart are denoted by Δ_Θ, Δ_β, Δ_Ξ, and their respective thresholds Ψ_Θ, Ψ_β, Ψ_Ξ. The counting variables are $FN^\wp$, $FP^\wp$, $h^\wp$, representing the accumulated number of false negatives, false positives, and hits for the current person $\wp$. They are used to compute the confidence ζ from the accumulated detection errors $\varepsilon^\wp$ and the hits $h^\wp$, and decide whether the current detected region has to be stored or discarded through the threshold Ψ_ζ.

reducing the number of both false positives and false negatives.

First, hands are manually annotated at the starting frames of each session to perform posterior color segmentation for the rest of frames. In this way, a GMM is learned with the marked set of most significative pixels, defining the skin color model of the person at those starting frames. Then, posterior frames are tested into the built GMM using a threshold ϑ, discriminating those pixels belonging to the skin color than those belonging to the background. The resulting blobs

are filtered using mathematical morphology closing operation with a 3 times 3 square structured element to discard noise and get smoother regions. Once the set of blobs is obtained, we need to choose those two candidates that belong to the hand regions. This is firstly performed by computing the optical flow among consecutive frames, which allows to discard noise on cases that we obtain more than two blobs by keeping those with higher movement.

On the other hand, we use the same heuristic procedure (Figure 3) than the one applied to the face analysis step for choosing the two best candidates, but now adapted to the hand regions. The incorrect regions detected at the first instance are those blobs having the highest optical flow, and then the heuristic procedure corrects those regions by comparing them with the hand regions obtained from the previous frame. As it occurs 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 manual annotation of the hand regions for the set of frames where 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.

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 [20].

2.4.3 Upper Body Analysis

As presented above, the probability of each pixel of an image to belong to a labeled body part is computed in section 2.3 using depth features. This information is used for posterior calculation of optical flow on RGB images where the upper body region appears. Therefore, each pixel of the image detected by Random Forest with high probability of being part of the person is used to calculate the optical flow. Finally, an average of optical flow is computed for the upper body region. Averaged optical flow will be used later to define behavioral indicators extracted from the upper body.

2.5 Behavioral Indicators

Once multi-modal features have been extracted, we use them to build a set of behavioral indicators that reveal communicative cues about each part involved in the conversation. Thus, in this section we describe the set of behavioral indicators constructed from the output of the different blocks shown in Figure 1, which will define the final feature vector for each part within the conversation.

2.5.1 Target Gazes Codification

Values regarding to the head angle pose of participants are correlated among them depending on the session setup. From them, one can be able to identify if participants are looking at other participants (i.e. target gazes). Since each conversation session have different number of people, target gazes have to be correlated depending on the number of people and their roles within the session (see section 3.1 for details about data). This correlation is performed by assigning binary values that codify target gazes through angle ranges. These angle ranges limit the people view areas for each participant role within a session and vary depending on the session setup. Then, a head pose angle that falls within these ranges means that the person (belonging to a role) is looking at the target subject (who may belong either to another or to the same role). Finally, we use these binary values to compute the time percentages of target gazes for each part.

2.5.2 Agitation Estimation

As explained in section 2.4.2, the positions belonging to the hand regions are computed from the extreme positions of higher optical flow. From them, we are able to quantify the movement for each region among consecutive frames. Therefore, to compute the accumulated agitation of the hands we calculate the averaged Euclidean distances of both hand positions among λ consecutive frames.

On the other hand, in section 2.4.3 is explained how the averaged optical flow is obtained for the upper body region. Therefore, to compute the accumulated agitation of the upper body we calculate the average of optical flow produced by the upper body among λ consecutive frames.

Thus, for each part and session, agitation averages are computed over processed frames, having a total of 8 agitation indicators, both alone and combined with other indicators previously calculated. These indicators regard to several possible combinations taking into account both agitation from the upper body and hands, and that agitation while a target gaze exists.

2.5.3 Posture Identification

From the $3D$ body position, we detect the body posture as a behavioral indicator, which may describe the involvement (or engagement) of the participants within the 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 done for agitation estimation, using Euclidean distances computed over $3D$ positions.

2.5.4 Speech Turns/Interruptions Detection

The speech diarization process of section 2.2.1 detects time segments belonging to each participant within the conversation. In order to extract the degree of interaction, we not only use the time each participant is speaking, but we extract the number of turn taking in each session. It allows to differentiate between a session where each part exposes its position from sessions and the different persons involved in the process.

Apart from the quantization of the turn taking, a relevant indication on the social communication analysis is the detection of interruptions, which are related to the dominance and respect between two persons [22]. Using the time between turns, we compute the percentage of the turns where a participant interrupts another participant.

2.6 Classification

The total number of behavioral indicators is 34, which defines the feature vector for each sample of our data set. Here, we define a sample as each part involved in a session keeping out an specific part (the mediator role). Thus, a

sample of 34 features is created for each (non-mediator) part and session. Note that each part may consist of more than one person. Table 1 summarizes the behavioral indicator list with their brief descriptions. All features except the features from f_7 to f_{12} are automatically obtained. These features [f_7,f_{12}] are extracted from a report obtained before the session, which is written by the mediators to analyze if they have influence on the final agreement prediction.

On the other hand, observations or responses that we want to predict in the classification task are the accuracy when correlating the agreement produced among the parts with the impressions given by the experts (mediators in our case). These opinions given by the experts are quantified values of agreement from 1 (lowest) to 5 (highest), assigned to each session based on level of agreement achieved at the end of the process by the two parts (victim and offender). Then, the ground truth of the system is obtained from a binarization of these values, which are assigned as the labels to each sample of the data set to obtain a binary setup. Since agreement is globally assigned for each session, those sessions containing two parts (or roles) will share the same ground truth labels for both generated samples. Section 3.1 explains in more detail the binarization task to set the sample labels.

Table 1: Summary of behavioral indicators defining each feature vector.

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

Learning is then performed over these samples and their features as a binary classification problem, grouping into two classes the quantified answers provided by the experts. For this, we base on four classical techniques regarding to machine learning field: Adaboost [23], Support Vector Machines (SVM) [24], and two kinds of Artificial Neural Networks (ANN), in particular Cascade-Forward (CF) and Feed-Forward neural networks (FF) [25].

3. EXPERIMENTS

This section presents the experiments performed by using the behavioral indicators summarized in Table 1. First, we describe the setting and validation measurements, as well as the performed experiments.

3.1 Data and Settings

Data consist of a total of 26 recorded conversational sessions of about 35 minutes-per-session. Each session contains audio-RGB-depth information, whose modalities are registered using the parameters from the camera, and synchronized among the different devices through the system clock. The set of images for each session has been recorded at resolution 640×480 and 12 frames per second (fps) in average, both for RGB and depth information. Each audio channel belonging to one of the four microphones linearly spread out along a multi-array microphone processes 16-bit audio at a sampling rate of 16 kHz. The distance between participants and the Kinect^{TM} device is between one and two meters depending on the limitations of recording facility. We defined the ground truth of the system by using quantitative answers in terms of agreement, provided by the mediator surveys. The agreement is defined as the degree of accordance produced among the parts (globally quantified for each session). Therefore, each part of a video session is a sample for the classification task, and the total number of used samples is 28.

Learning is performed using leave-one-out validation, keeping each time one sample out for testing. Since the total number of samples is reduced and the ground truth values are quantified within ranges $[1, 5]$, we simplified the problem by grouping the different response degrees into binary groups. As we are defining a binary setup, then the value 3 can be considered either as high or low. For this reason, we have thrown the experiments twice considering both cases and computing the mean of the two leave one out experiments.

In order to accomplish the compromise described at the end of section 2.4.1, the parameters used in the heuristic procedure have been 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.

In our experiments, we have set the standard value of 50 to the number of decision stumps for Adaboost technique. For the SVM, we have experimentally set to 1 the cost parameter, and 0.5 the gamma parameter to the radial basis function. Finally, we have set two standard neural networks (CF and FF), 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.

3.2 Results and discussion

The addressed predictions for our classification task focus on the reached agreement level. The percentage of accuracy on predictions is then compared among the different techniques: Adaboost, SVM, CF, and FF. Results for predicting the agreement are shown in Table 2. The best prediction is given by the FF neural networks with a 75% of accuracy. Moreover, we can note that all classifiers are able to predict above the random decision. This fact can be interpreted in the way that there exist a correlation degree between the captured data and the information that we want to predict.

Table 2: Accuracy predicting agreement.

Label	Adaboost	CF	FF	SVM
Agreement	71%	71%	**75%**	71%

An important aspect to highlight in the classification task is the weight of grouping the quantified degrees of mediator answers into the binary case. This entails to obtain different results depending on the employed classification technique and the prediction type. This is probably due to the uncertainty of the mediator when assigning a value of 3 to the answers with respect to the evaluation purposes, fact that may include noise to the overall data.

As described in section 2.4.1, the user manual interactions is an important requirement in our proposed semi-automatic system to improve the continuity of positive detections among consecutive frames. For our sessions, the averaged frequency rate of manual annotations required is 1 for each 2000 frames using the above parameters. It means that using those parameters the feature extraction procedures for hands and faces offer high accuracy.

4. CONCLUSION

We proposed a multi-modal framework for the analysis of non-verbal communication in conversation settings. We showed the usability of computer vision, signal processing, and machine learning strategies in these scenarios. In particular, we computed a set of features from audio-RGB-depth data. Then, a heuristic procedure was presented within the multi-modal feature extraction to improve the continuity of positive detections among consecutive frames. Finally, we defined an automatic computation of behavioral indicators used as final features for learning and classification tasks. We demonstrated the applicability as a tool for the experts, obtaining results upon 75% of accuracy predicting the agreement in conversational victim-offender mediation processes based on the ground truth defined by the experts. Based on the obtained results, our future work involves including local behavioral features, which will provide information about the instant of time where the behavior takes place (early or latest stages of the conversational session). Additionally, we plan to extend the binary agreement classification problem to a continuous or regression task, where a more fine agreement prediction could be achieved.

5. REFERENCES

[1] A.-S. Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, Massachusetts, 2008.

[2] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, and D. Gatica-Perez. Body comunicative cue extraction for conversational analysis. *FG*, 2013.

[3] 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*, 7(1-2):39–53, 2013.

[4] A. Vinciarelli, H. Salamin, and M. Pantic. Social signal processing: Understanding social actions through nonverbal behaviour analysis. In *CVPR*, volume 3, pages 42–49, 2009.

[5] S. Escalera, O. Pujol, P. Radeva, J. Vitrià, and M. Teresa Anguera. Automatic detection of dominance and expected interest. *EURASIP Advances in Signal Processing, Research Article*, 2010.

[6] T. Takahashi and F. Kishino. Hand gesture coding based on experiments using a hand gesture interface device. *SIGCHI Bull.*, 23(2):67–74, March 1991.

[7] G. Fang, W. Gao, and D. Zhao. Large-vocabulary continuous sign language recognition based on transition-movement models. *SMC-A*, 37(1):1–9, 2007.

[8] V. Ponce, M. Gorga, X. Baró, and S. Escalera. Human behavior analysis from video data using bag-of-gestures. In *Int. Joint Conf. on Artificial Intelligence*, volume 3, pages 2836–2837, 2011.

[9] J.-I. Biel and D. Gatica-Perez. Vlogsense: Conversational behavior and social attention in youtube. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7S(1):33:1–33:21, November 2011.

[10] D. McNeill. *Gesture and Thought*. University of Chicago Press, Chicago, 2005.

[11] A. Mehrabian. *Nonverbal communication*. Aldine-Atherton, 1972.

[12] A. Hernández-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, S. Escalera, M. S. Kamel, and F. Karray. Human limb segmentation in depth maps based on spatio-temporal graph-cuts optimization. *JAISE*, 4:535–546, 2012.

[13] A. Hernandez-Vela, M.-A. 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. *International Conference on Pattern Recognition*, pages 449–452, 2012.

[14] 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 rgb-d data. *ICPR*, 2012.

[15] H.-D. Yang, S. Sclaroff, and S.-W. Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE PAMI*, 31(7):1264–1277, 2009.

[16] A. Stefan, V. Athitsos, J. Alon, and S. Sclaroff. Translation and scale-invariant gesture recognition in complex scenes. In *PErvasive Technologies Related to Assistive Environments*, pages 7:1–7:8, 2008.

[17] 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 *in Interspeech*, 2005.

[18] X. Anguera and J.-M. Pardo. Robust speaker diarization for meetings: Icsi rt06s evaluation system. In *in Proc. ICSLP*. Springer Verlag, 2006.

[19] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[20] M. Fisher. Interpreting sensor values.

[21] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. *CVPR*, 2012.

[22] S. Escalera, X. Baró, J. Vitrià, P. Radeva, and B. Raducanu. Social network extraction and analysis based on multimodal dyadic interaction. *Sensors*, 12(2):1702–1719, 2012.

[23] Y. Freund and R.-E. Schapire. Experiments with a new boosting algorithm. In *In Proceedings of the 13th International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1996.

[24] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[25] M. Gopikrishnan and T. Santhanam. Effect of different neural networks on the accuracy in iris patterns recognition. In *Int. Journal of Reviews in Computing*, pages 22–28 vol.7, 2011.