Body communicative cue extraction for conversational analysis

Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, Laurent Nguyen and Daniel Gatica-Perez

***Abstract*— Nonverbal communication plays an important role in many aspects of our lives, such as in job interviews, where *vis-à-vis* conversations take place. This paper proposes a method to automatically detect body communicative cues by using video sequences of the upper body of individuals in a conversational context. To our knowledge, our work brings novelty by explicitly addressing the recognition of visual activity in a seated, conversational setting from monocular video, compared to most existing work in video-based motion capture, which targets full-body with lower limb activities. We first detect the person hands in the sequence by searching for the higher speed parts along the whole video. Then, aided by training a set of typical conversational movements, we infer the approximate 3D upper body pose, that we transfer to a low-dimensionality space in order to perform action recognition. We test our system in the context of job interviews, with several new databases that we make publicly available.**

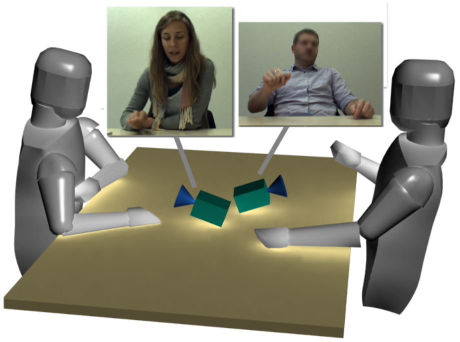


Fig. 1. Our proposed framework outputs hand position, speed, approximate upper body 3D pose, and estimated ongoing activity, using single camera conversational video sequences as input.

I. Introduction

Nonverbal communication plays a significant role in how we perceive each other in a social context [14], [24]. Some key aspects of life, such as job interviews, take place in *vis-à-vis* conversations, in which the way we are socially perceived has a significant weight in our success [5]. This subject has therefore been intensively analyzed in social psychology and cognitive science [14].

However, there has always been the need for an interpreter. That is, a person that emits a judgement on the perceived traits of the analyzed subject, or that codes specific behaviors. In order to address this problem, we propose a new method to analyze, in an automatic way, upper body nonverbal cues of people in a conversational context. By using frontal videos of a person discussing around a table as input, we developed a set of new computer vision algorithms in order to first extract, and then obtain a series of measurements that allow to analyze upper body movements and actions of a person with conversational meaning (see Figure 1).

A significant amount of research [30] has been done in order to obtain body movements, information generally known as markerless motion capture. Moreover, this topic has been identified to be a hard problem to solve when using a single camera system. This paper proposes a novel upper-body motion capture system based on multiple image cues, such as face, hand and motion detectors, and using training data from 2.5D labeled conversational video sequences. As a result of our work, we are able to determine the 3D upper body pose of a person and its corresponding hand speed and a number of conversational actions, which is the first building block for more advanced interaction analysis techniques. Figure 2 shows a general diagram of the framework.

A. Related work

1) Computational modeling of interaction: As discussed in [11], a significant amount of literature on analysis of nonverbal social interactions has been published over the years. Interactions among small groups and dyads have been studied. Much work has investigated basic features (e.g. visual motion or basic hand gestures) that can be robustly extracted from video, but that correspond to rough representations of actual activity [26]. The subject has also been approached by the wearable computing field, where people wear sensors to be able to capture body motion and posture in conversations [7]. But while motion can be accurate, there is a need to place and wear intrusive devices. There is also a substantial amount of work in hand gesture recognition [21], [22], which in general has had different emphases (i.e., human-computer interaction or sign language recognition) than the one we address here.

Action recognition with computer vision can be applied to automatically obtain body communication cues. The most traditional approach in these systems is to first get the body pose and then analyze it [10]. Most recent works are able to do this without getting the body pose (i.e. without performing motion capture) through diverse techniques [25]. However, it has been shown in practice [2] that, even though it is possible to perform activity recognition without knowing the body pose, it is still beneficial to know it. In this work we are looking for specific nonverbal cues such as adaptors, which are movements like head scratching, that provide information about attitude, anxiety level and self-confidence [18]; and beat gestures, which are flicks of hands used to emphasize

{amarcos, marta, pizarro}@depeca.uah.es, University of Alcala, Spain
{gatica, lnguyen}@idiap.ch, Idiap Research Insitute, EPFL, Switzerland

important parts of the speech with respect to the larger discourse [19]. Body posture is also found to be an important indicator to the emotional state of a person [20].

2) Motion capture: Motion capture has been a long-standing subject in vision and graphics, as it allows for a number of interesting applications, mainly in the fields of virtual-character animation and gaming [31]. Traditionally, there was a need for wearing cumbersome sensors in the body, compromising the practical aspects of the concept. In recent times however, markerless motion capture solutions heavily removed the need for them, by being able to obtain the body pose with sensors in the environment (i.e. cameras of different kind), with relatively high precision. This is why this method in particular has generated a lot of literature in recent years [30]. There are several approaches, which can be grouped into single-camera (monocular) systems, multi-camera systems, and range camera systems.

Given that the human body movement is intrinsically non-linear and high-dimensional [31], having several point of view of the body helps to remove ambiguities. Although a lot of alternatives exist, the traditional approach when having multiple cameras, is to first obtain multi-view silhouettes of the body and then iteratively adapt a model to these silhouettes [34]. However, settings with more than one camera are not always available.

Recently, the popular Kinect device [29] has spread markerless motion capture systems based on range cameras (i.e. so-called 2.5D images), which provide a depth measure for every pixel, removing many of the problems of regular camera systems. Kinect features a body part detector implemented by training a random forest with a huge number of synthetically-generated poses, and since recently, is also able to track upper-body only. Related to this, a few works have arisen [17], [9]. However, [17] needs to get a reference pose to initialize, and in [9], a camera calibration is needed. The main downside, however, is that this kind of sensors are not yet as extended as to be always an available option, especially in already recorded footage, which is the case for most work in social psychology and communication.

This point highlights the relevance of the monocular approach. A lot of vast and diverse work has been made here too [23], [6], [27] [8]. To address the problem in the conversational video sequences, we propose a robust cue such as dense TV-L1 optical flow [4], that together with a face detector, is used to first place the hands, and then with the help of a set of trained 3D movements, obtain an approximate measure of the human body.

Finally, as extense as the literature of markerless motion capture and multimodal interaction is, to our knowledge, a joint approach that explores the use and implications of having automatic body communication extraction in a social conversational context is still missing, although some works have been done in that direction, like [16], which analyzes gender, age, or animic state with gait cycles. The proposed work here presented is designed to be a first solution to fill this void.

B. Paper contributions and organization

Taking into account the related works previously reviewed, the main contributions of this work are:

1) A new method for extracting hand position from conversational video sequences, by exploiting the fact that optical flow is a strong indicator of where the hands are in conversation.

2) A new method for visual tracking, if the whole sequence is available from the start (typically the case in psychology, management and cognitive science experiments).

3) A new method for extracting 3D torso pose from 2D images in a seated person setting for action recognition.

4) An objective evaluation of the above tasks using a job interview dataset. This generated three public datasets. Two are used to evaluate hand position and activity recognition accuracy. One contains a set of 3D poses labeled with the help of a range camera.

The rest of the paper is organized as follows: in section II we present the proposed method for hand mapping, hand tracking and estimated torso pose retrieval algorithms; in section III we test them by using several self-made databases, and discuss our framework limitations, and finally in section IV we expose our conclusions.

II. Proposed method

We propose to build a set of modules for analyzing nonverbal cues, from a video of the upper body of a person. In order to accomplish that, we developed hand and face detectors that together with offline training data, allows to get the approximate 3D upper body pose. From this information we extract the conversationally relevant cues. See Figure 2.

A. Hand likelihood maps

Given a video frame I, where $I(\mathbf{p},t)$ is a pixel color at position $\mathbf{p} = (u,v)^\top$ and time index t, the goal is to obtain a measure of where the hands are in that image. In order to accomplish the goal, the features used should be as color/appearance invariant as possible, to increase the robustness, while also exploiting the constraints of the face-to-face interactive setting.

We hypothesized that, given a frontal, static camera pointing to the upper body of a person, hands are normally the parts of the image that show more movement. Two strong indicators are: a) they are usually the closest part to the camera, and b) they are the further body part from the body's axis of rotation, so they show the highest spatial speed for a given joint angular speed.

With this in mind, we built a 2D hand likelihood map, where high values mean that the expectancy of a hand being in that region is high. The hand likelihood map follows the assumption that: in an image, the hands are the skin-colored parts that are not the face and which show more amount of movement. In order to enforce that rule, we need to compute the optical flow of the sequence to extract movement information, skin segmentation, and face detection. Also, given the natural appearance of the fingers,

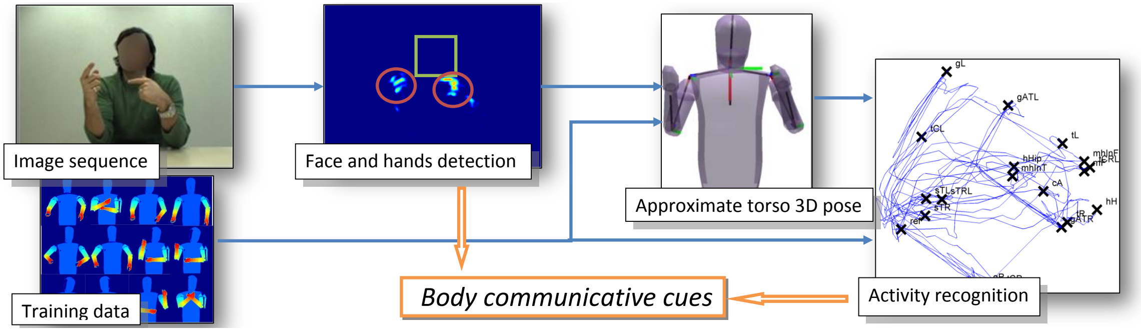


Fig. 2. Proposed framework to first extract and then analyze the body posture in conversational sequences.

which have lots of edges, we also used image edge detection as a feature. We detail the steps in Figure 3.

1) Video movement retrieval: In order to retrieve a measure of the characteristics of the movement in the image, we use a state of the art framework [4], that provides smooth optical flow (see Figure 3) by performing convex optimizations while allowing for outliers. It outputs the optical flow modulus I_{OF}, that can be seen as a scalar image $I_{OF}(\mathbf{p},t)$.

2) Face detection and edge extraction: In order to detect the face of the conversating person in the video, we used a yet unpublished probabilistic version of the Viola & Jones face detector [32]. It uses likelihood information from the output of every adaboost cascade classifier, so that the output is probabilistic rather than binary. We built a face mask, set to 0 inside the face region of interest and 1 otherwise, so that later the face pixels are not taken into account when computing I_F. We use a simple Canny edge detector [3] with a low threshold, to obtain an edge map which we then smooth in order to better search for maxima in the hand likelihood map. We get the final edge map I_E.

3) Skin segmentation: Inspired by [28], we profited from the face detection to infer the skin color of a given person, as the skin color of the face and hands are usually similar. A number of N_{RF} sequence frames are randomly picked. It is established by [12] that skin color hues usually fall within the $(0,0.2)$ range of the hue channel in an HSV image. We therefore built skin color statistics (μ_{SH},σ_{SH}) using the image pixels of the N_{RF} detected face regions that fell within that range. A binary segmentation is then performed, in which a pixel is labeled as skin if it falls within a distance σ_{SH} of the skin hue mean μ_{SH}. We tried the algorithm with people of different skin tones (Figure 3). Simple per-pixel morphological operations are then performed with the segmented image, in order to obtain the final I_S binarized result image.

4) Hand likelihood map formation: The hand likelihood map is obtained as the intersection $I_H = I_{OF} \cdot I_S \cdot I_F \cdot I_E$ of all these cues, as their simultaneous verification constitutes the set of rules that we established in II-A it should follow. See Figure 3 and the **supplementary video material**.

B. Hand tracking

Exploiting the fact that the whole video is available since the beginning of the processing phase (something that is typical of offline settings), we aim to detect the hands by performing an analysis of the obtained hand likelihood map sequence in order to model the premise of the hands being the quickest part of the human upper body along a sequence. A tracking scheme is implemented in two steps, as follows.

1) Hand likelihood map clustering and tracklet extraction: For each hand likelihood map frame I_H, we first perform a search for local maxima, first by using a smoothing filter in order to better show local tendencies, obtaining I_{HS}. We then threshold I_{HS}, and cluster local the obtained maxima using an adaptative k-means classifier with support for a variable, unspecified number of classes, which also provides identity consistency of the several local clusters along time. At this point we have a set of local maxima of the whole sequence of hand likelihood maps.

We extract the paths of those several local maxima detected in I_{HS}. These set of N_t hand likelihood map trajectories, that we call tracklets, are non continuous in the sense that detected local maxima will disappear and then re-appear in the image because of occlusions, being out of frame, and/or malfunction of the hand likelihood maps. The output of this is a group of tracklets, each of which is defined as follows:

$$\{\mathscr{T}_i\}_{i=1}^{N_t} = \{t_{0,i}, t_{f,i}, \lambda_i, \mathbf{p}_{0,i}, \mathbf{p}_{f,i}\}_{i=1}^{N_t} \qquad (1)$$

where N_t is the number of tracklets in the sequence; $t_{0,i}, t_{f,i}$ are the time instants when the tracklet i starts and ends; λ_i is the accumulated likelihood along the tracklet i duration. Longer tracklets therefore usually have bigger λ_i. As tracklets do not have a maximum length value, λ_i is not upper-bound; $\mathbf{p}_{0,i}$ is the pixel position where the tracklet i started, and $\mathbf{p}_{f,i}$ is the pixel position where the tracklet i ended.

2) Finding hands' most likely paths: In order to obtain the best 2D paths for a hand in the image, we implement a decision tree algorithm, in which the tracklets are the branches, and a decision of what tracklet to follow next is made in every node, based on several factors explained below. In Figure 4 a 1D example of how four tracklets look

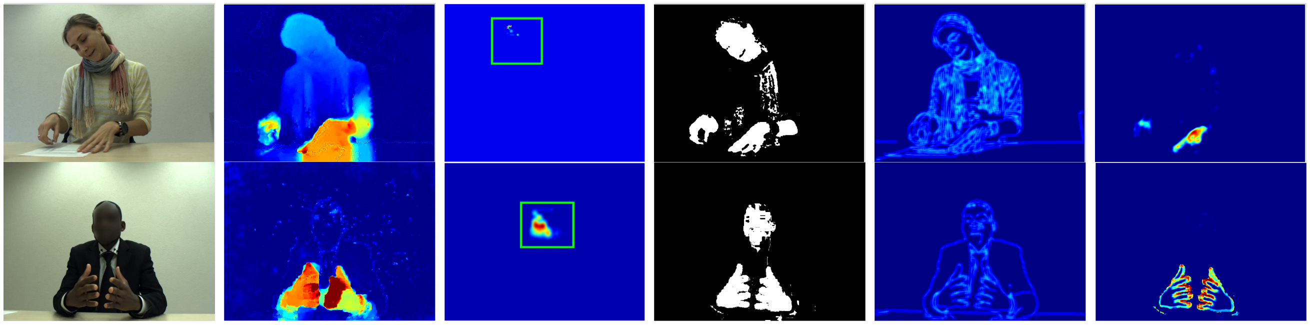


Fig. 3. Steps for building the hand likelikihood maps. All images in the same row corresponds to the same time instant (see the supplementary video for a display in motion). Columns, from left to right: input video frame, optical flow normalized modulus, probabilistic Viola & Jones output, skin segmentation and face ROI, edge map, and hand likelihood map (the intersection of the other cues). Best viewed in color.

along time is shown. The goal is to find the path in which the accumulated likelihood is maximum. For this, we establish three basic rules:

- Once the hand is assigned to a tracklet, it is not possible to jump to another tracklet until the current one has reached its end.

- Once a tracklet has finished, it is possible for the hand to stay in that tracklet final pose until the end of the sequence, or jump to any other tracklet that has started afterwards.

- When jumping from one tracklet to another, jump distances (in pixel positions) are taken into account to penalize far jumps. The accumulated likelihood of a hand taking two tracklets, $\mathscr{T}_i$ then $\mathscr{T}_j$ (that is, following path from the initial point of $\mathscr{T}_i$ $\mathbf{p}_{0,i}$ to the final point of $\mathscr{T}_j$ $\mathbf{p}_{f,j}$ through points $\mathbf{p}_{f,i}$ and $\mathbf{p}_{0,j}$), separated by a distance $d_{i,j} = \|\mathbf{p}_{i,f} - \mathbf{p}_{j,0}\|$, is:

$$\Lambda_{i,j} = \lambda_i + e^{\rho(-d_{i,j})}\lambda_j \tag{2}$$

Where ρ is a distance penalization factor (manually set in experiments). We process the existing paths, and then get the one which contains the highest accumulated likelihood. The sequences in consideration in our work are long (up to 20 minutes), and the number of tracklets could be in the hundreds. Given that the number of paths increase exponentially with the number of nodes, we back-compute the accumulated likelihoods, retaining only the maximum path at each node.

Taking both hands into account: The goal is to have the two highest likelihood hand trajectories, given that there are two hands to track. Therefore we search for the two best paths along the tracklet tree. In order to do this, we define a priority hand, that is, the one that will evaluate the tracklet tree first, thus getting the best path.

After this has been computed, we set to 0 the accumulated likelihood of the tracklets used by the optimal path, and then evaluate the modified tree for the other hand. This algorithm finally outputs the position of the visible moving hands $H_1(t) = (\mathbf{p}_{H1})$ and $H_2(t) = (\mathbf{p}_{H2})$ in the image at time t.

C. *Torso pose extraction*

In order to infer the torso 3D pose, we propose to use the 2D hand and face position, together with training data, which allows to map the 2D observations in the image into a 3D estimated pose. To do this, we first collect and label several typical and conversationally relevant upper body poses, with the help of a range camera. Then we create a series of synthetic observations with the collected poses to compare them to the real ones. The process is explained as follows.

1) 3D torso model: We use a synthetic 3D polygonal torso model, driven by an underlying N_j-joint skeleton (see Figure 5), whose pose is parameterized by the 3D euclidean rotation angles $\Phi = \{\alpha_j, \beta_j, \gamma_j\}_{j=1}^{N_j}$ of every body segment, relative to the root node (the base of the neck joint), which is referenced to the world global coordinates by its 3D position and orientation $\Psi = \{\alpha_R, \beta_R, \gamma_R, x_R, y_R, z_R\}$.

We have not experienced any problems regarding Gimbal locks, therefore we did not deem necessary the usage of alternative angle representations such as quaternions ([15]).

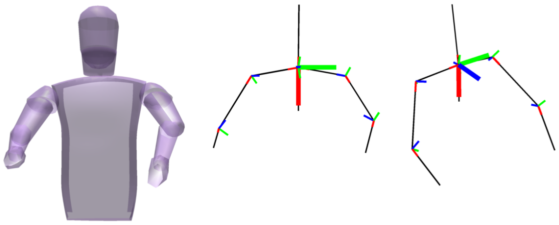
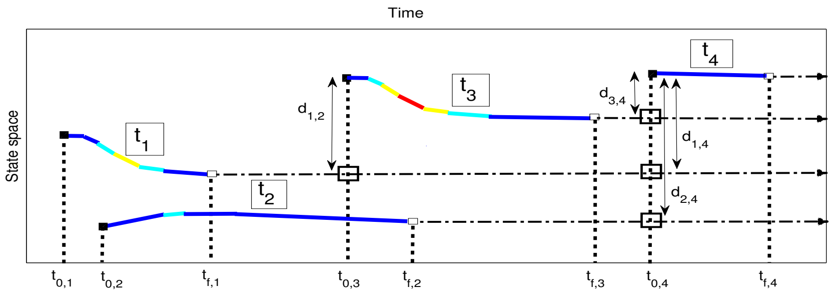


Fig. 5. Torso Model. Left: 3D mesh. Centre and right: underlying skeleton model. The base of the neck is the root node.

2) Training process: The training process is what gives the ability to infer the 3D articulated model of a torso from 2D observations. It is only done once, as a pre-processing step, it should therefore not be mistaken for user assisted methods. Two subjects (two male, two female) are recorded with a range camera in a similar setting to the target scenario (i.e. sitted by a table, see Figure 7 left), while performing a set of N_{Ta} actions, resulting in a total of N_T training frames. See Table I below for the list of actions included in the training data set.

As can be seen, the above actions are all typical of a conversational setting. We first group these actions into a set



$$\Lambda_1 = \lambda_1$$
$$\Lambda_2 = \lambda_2$$
$$\Lambda_3 = \lambda_3$$
$$\Lambda_4 = \lambda_4$$
$$\Lambda_{1,3} = \lambda_1 + \lambda_3 e^{-d_{1,3}\rho}$$
$$\Lambda_{1,4} = \lambda_1 + \lambda_4 e^{-d_{1,4}\rho}$$
$$\Lambda_{2,4} = \lambda_2 + \lambda_4 e^{-d_{2,4}\rho}$$
$$\Lambda_{3,4} = \lambda_3 + \lambda_4 e^{-d_{3,4}\rho}$$
$$\Lambda_{1,3,4} = \lambda_1 + \lambda_3 e^{-d_{1,3}\rho} + \lambda_4 e^{-d_{1,4}\rho}$$

Fig. 4. Left: Hand tracking tracklet decision tree example with 4 tracklets ($\mathcal{T}_1$ - $\mathcal{T}_4$) and 4 nodes, along a 1D state space. Color encodes tracklet likelihood in a given time instant (warmer means higher). Nodes are represented with squares. Right: Likelihood values Λ for each possible path. Best viewed in color.

Label	Description
refPose	Reference pose, hands separated, on table.
cA	Arms crossed.
gR, gL, gRL	Perform conversational gestures with one hand, then with the other hand, then with both.
gTR, gTL, gTRL	Same as previous, but resting the non-used elbow on the table.
hH	Hands touching the back of the head.
hHip	Hands touching the hips.
sTR, sTL, sTRL	Placing one hand, then the other, then both in different parts of the table.
thkr	One hand one the chin, another one supporting the elbow of the hand that touches the chin.
tCR, tCL, tCRL	Touching the chin with one hand, then the other, then both, with the non-used hand resting on the table.

TABLE I

LABELED ACTIONS IN THE TRAINING PROCESS.

of 6 conversationally relevant categories, the four shown in Table II and two additional classes for 'handsOnHead' and 'joinHands'. Upon inspection of the data, we realized that these categories, while natural, occurred only in a negligible fraction of our data (approximately 0.5 % of total), and therefore are not considered for experiments. This grouping of subactions into four categories is done both to avoid increasing the variance of the error, and because this work being a first approach to the problem, which will be extended afterwards.

Category	Actions in category
hiddenHands	No hands detected
gestures	cA, gR, gL, gRL, gTR, gTL, gTRL, hHip, hH
handsOnTable	sTR, sTL, sTRL, refPose
selfTouch	thkr, tCR, tCL, tCRL

TABLE II

CATEGORIES OF ACTIONS

The range camera provides a set of (u_{RC}, v_{RC}, z_{RC}) observations, where (u_{RC}, v_{RC}) are the usual 2D image coordinates, and z_{RC} is the depth value. Manually annotating the position of every joint in the (u_{RC}, v_{RC}) space of the depth recordings allows to obtain the 3D location $\{J_{T,k}\}_{k=1}^{N_T} = \{x_{Tjk}, y_{Tjk}, z_{Tjk}\}_{k=1,j=1}^{N_T,N_j}$ of every joint, along the different training sequences, where N_j denotes the number of joints. If an occlusion occurs, we estimate the position of the occluded joint either as the same one of the last frame, or as an estimated guess. Given that our torso model is parameterized by angles, and at this point we have a set of 3D points J_T, we use an optimization fitting scheme, by using non-linear least-squares to get the angles $\{\Phi_{T,k}\}_{k=1}^{N_T}$ from the 3D points J_T. Given that more than one combination of angles could result in the same 3D joint positions, we establish an angle limit for every joint, and then build an energy function based on these constraints, so that the energy is minimum the further away the joint is from the limit (see Figure 6).

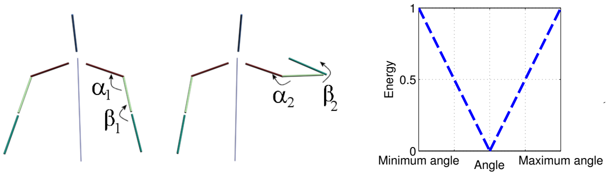


Fig. 6. Left skeleton: low energy arm pose. Right skeleton: high energy arm pose (given that α_2 and β_2 are closer to the maximum angle than α_1 and β_1). Right graph: energy function.

Even if rough, this setting gives good results in obtaining the desired parameterization (see Figure 7, and 3D mesh in Figure 5, which shows natural limbs and head orientation, thanks to the built joint angle energy function).

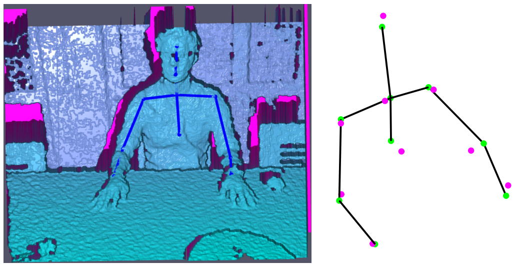


Fig. 7. Left: manually annotated 3D skeleton overlaid into the range camera 3D measurements (see supplementary video). Right: optimized torso pose relative to the manually annotated points (in magenta). Best viewed in color.

We then input the joint angle information of these actions using the Principal Component Analysis (PCA) framework [13], into a low-dimensional latent space:

$$\{\Phi_{T,k}\}_{k=1}^{N_T} \mapsto \Pi_{T,k} = \{\pi_{1,k}, \pi_{2,k}, ..., \pi_{N,k}\}_{k=1}^{N_T} \qquad (3)$$

We are aware that there exist more modern alternatives to PCA such as [33]. However, as their study is outside the aim of this work, we rely on the simple and efficient two-way mapping that PCA provides.

After the movement is compressed by PCA, we manually mark the most characteristic instant of every action in the PCA low-dimensional latent space (we call them key points, $\{K_{Ti}\}_{i=1}^{N_{Ta}}$). This is because along a movement sequence, there are intermediate instants in which the main characteristics of the final posture are not captured (see Figure 8).

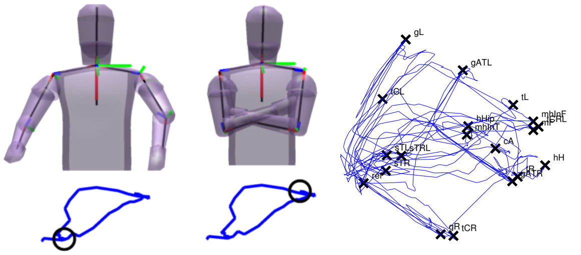


Fig. 8. Left and center: two poses of the cross arms action, with their representation in the PCA space. Only the center one is characteristic of the action (the key point). Right: Whole set of trained motions in the latent space (see Table I), with the labeled key points of every action (black crosses).

3) Training data integration: After the process just described, we construct, for every training frame, synthetic observations for the hand and face positions, projecting them from our torso model onto 2D images by using an estimation of the camera extrinsic parameters. Afterwards we compare the real inputs with our set of synthetic data, by using discrepancies between hand position and foreground/edges to choose the best match.

The angle data of the noisy set of training poses are then smoothed with a Kalman filter, to obtain the final output pose sequence $\Omega_t = \{\Phi_t, \Psi_t\}_{t=1}^{N_{frames}}$.

D. Feature extraction and conversational cue inference

At this point we have obtained the hands' position in the image $\{H_1(t), H_2(t)\}_{t=1}^{N_{frames}}$ and the approximate 3D torso pose $\{\Omega_t\}_{t=1}^{N_{frames}}$. In order to obtain conversationally relevant information, we extract the following features:

- *Average hand speed:* Given that we have the positions of the two hands along each sequence (Section II-B), we compute their speed to get the average speed $H'_{1,2}$ along a sequence.

- *Action recognition histograms:* In order to infer the action that is being performed at a given instant of the sequence, we first get the pose point in the latent space by computing the mapping $\Phi_t \mapsto \Omega_t$ (Section II-C.2, Eq. 3), and then compute the Euclidean distances in the latent space to every key point K_T, which we use to compute the winner category (4 in our case). This outputs a label with the estimated performed action for every sequence, that we use to build an action histogram, which we also use as a feature, because it provides a measure of how a person has moved along the sequence. Figure 9 illustrates the extracted features for two people with significantly different amount of hand motion and body postures.

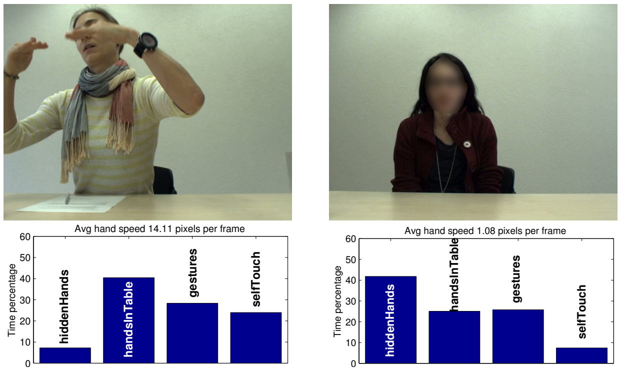


Fig. 9. Extracted features of two different persons. See supplementary video.

III. Implementation and Results

A. Implementation and data

We found a lack of appropriate public databases for testing activity recognition and hand tracking in a compatible seated, conversational, long sequence setting. Therefore we built a set of experiments that we make public in [1], in order to test the performance of the hand tracking and action recognition algorithms (Figure 12).

We use data obtained from real job interviews, taking place in a conversation room using two uncalibrated but synchronized HD cameras (we resize the images to 640x480 as it is enough for our purposes) with a sampling rate of 26 frames per second, fixed on a table and pointing to the upper body of the participants (see supplementary material). This data consists of 8 full jobs interviews with 105 minutes of conversation sequences, in which 8 different subjects appear.

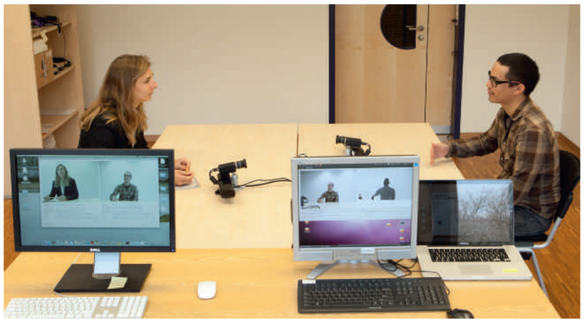


Fig. 10. Video recording setup.

In order to reduce the number of frames to process, and given that we use optical flow to detect the hands, we filter the segments of the video in which there is not enough image difference, by using a manually set threshold on the frame difference signal. That is, we do not process the frames which do not show enough change, although we take them into account when computing the average hand speed.

For the hand tracker, we manually labeled the position of the hands in a challenging 1450-frame sequence, where a person wears a skin-colored scarf and has no sleeves. It is therefore very useful to determine how well the proposed hand map behaves with the help of optical flow and edge information, in comparison to a regular skin segmentation. The error is measured in two ways: (1) on the image plane, in pixels, and (2) as a detection rate, that measures how often the number of detected hands (0, 1, or 2) is correct.

For testing the action recognition algorithm, we manually labeled the actions performed by the 8 different subjects, according to the categories in Table II. To simplify the process, we labeled one every 15 frames (or approximately 6 tenths of a second) in the portions of the video which showed movement above the manually set threshold. This resulted into 2590 manually labeled frames, see Table III for the split per category. As performance measure, we use frame classification accuracy.

hiddenHands	gestures	handsOnTable	selfTouch
2.23%	24.56%	67.53%	5.67%

TABLE III

CATEGORY FREQUENCY IN THE USED DATABASE.

B. Results and discussion

The results are shown in Figure 12 and illustrated in Figure 11 and the supplementary material. Regarding hand tracking, Figure 12 (top) shows the error for both hands. As can be seen, the error remains below 20 pixels in many frames, except when error spikes appear. The mean error is **17.35 pixels**. Furthermore, the detection error is **8.75%**. Note that the chosen data for testing is specially challenging, so we would expect the method to perform better in many other situations.

Regarding action recognition, the overall classification accuracy is **72.5%**. The performance is significantly better than random (25%), but also than a majority-class method that would label every frame as 'handsOnTable' (67.5%, $p = 0.0238$). It is important to mention that correctly classifying the 'handsOnTable' action is not trivial, as factors like slow movements, skin colored clothes, or sleeve-less shirts have to be dealt with. As an illustration, we show two failure examples of the hand tracking in Figure 13. Examples of correctly recognized actions are shown in Figure 11.

The algorithm finds its main challenges in two points: (1) Given that we make a comparison with training data in order to obtain the torso 3D pose, the system has difficulties coping with body poses outside the training ones. This can be addressed in two different ways: by creating a larger training set, where using synthetically generated poses is an option [29], or by using the current output to initialize an optimization scheme to better adjust the pose. The latter option could be viable only if the processing time is low enough, to keep the problem tractable given the large amount of data to process. (2) As we perform the analysis on monocular video, the observed hand position if the subject makes hand gestures in front of his face is very similar to that of self touch. Similary, judging exclusively the wrist joint position, it is challenging to differentiate between actions 'gestures' and 'handsOnTable', if the action is taking place near the table.

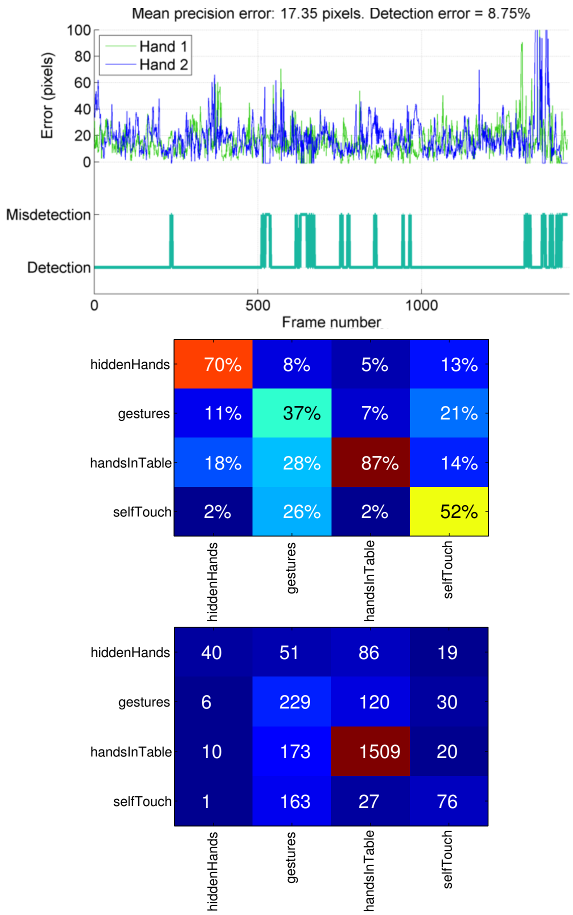


Fig. 12. Top: hand tracking error results. Middle and bottom: Confusion matrix, normalized by columns (middle) and not normalized (bottom). Warmer colors mean higher values. Best viewed in color.

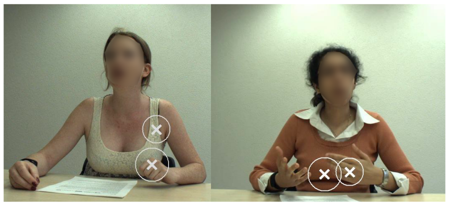


Fig. 13. Failure examples while the subject has the hands on the table.

IV. CONCLUSION

We present a system that automatically analyzes the communicative cues of seated participants in conversational events recorded with regular, broadly available cameras (although range cameras are needed only for the offline

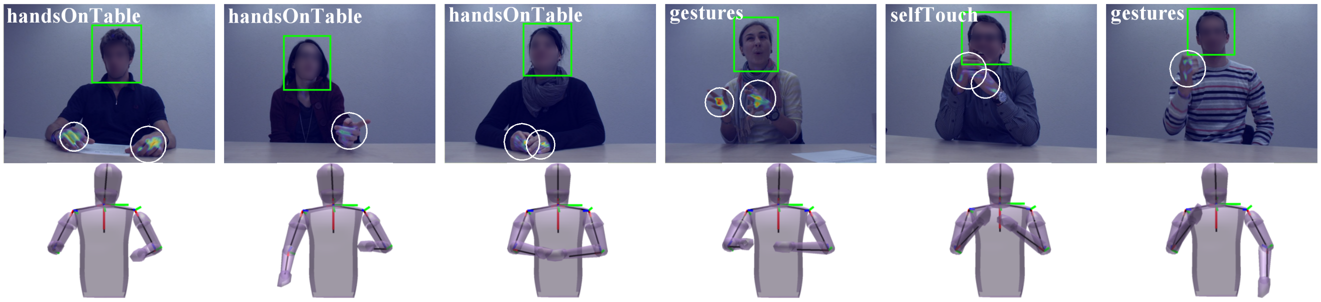


Fig. 11. Frame results. Top row: input frame, with hand likelihood map, face detection, hand tracking and recognized action overlapped. Bottom row: output 3D torso pose. See supplementary video.

training phase). We built original hand and face detectors which were used to get an approximate 3D upper body pose, which will be useful for developing more complex techniques in our future work. This pose, highly dimensional, is compressed to reduce its dimensionality and perform action recognition. With this information we propose to use average hand speed and action histograms as descriptors of body communicative cues. Specifically, we look for adaptors and beat gestures, which previous studies have shown to carry nonverbal communication information. Our system can recognize basic upper-body actions with an accuracy of 72.5%, in a dataset of 105 minutes of real job interviews.

The results obtained with our proposal have shown to be useful as a first building block to automatically analyze psychological traits of the participants in the conversation, and psychologists that we have discussed with find this type of recognition and current performance quite promising. Our future work will deepen and explore the possibilities that this fusion of disciplines give.

V. Acknowledgments

This research was funded by SNSF SONVB project, the Spanish Ministry of Science and Innovation under project VISNU (ref. TIN2009-08984) and the University of Alcala FPI internship program. We thank Denise Frauendorfer (Université de Neuchâtel) for the support with data collection.

References

[1] http://www.idiap.ch/project/sonvb/databases.

[2] G. F. Angela Yao, Juergen Gall and L. V. Gool. Does human action recognition benefit from pose estimation? In *BMVC*. BMVA Press, 2011.

[3] J. Canny. A computational approach to edge detection. *PAMI*, 1986.

[4] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems withapplications to imaging. *JMIV*, 2011.

[5] J. Curhan and A. Pentland. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first five minutes. *Journal of Applied Psychology*, 2007.

[6] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Articulated human pose estimation and search in (almost) unconstrained still images. Technical report, 2010.

[7] S. Feese, B. Arnrich, G. Troster, B. Meyer, and K. Jonas. Detecting posture mirroring in social interactions with wearable sensors. *Wearable Computers, IEEE International Symposium*, 2011.

[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[9] J. Gall, A. Fossati, and L. J. V. Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, 2011.

[10] J. Gall, A. Yao, and L. J. V. Gool. 2d action recognition serves 3d human pose estimation. In *ECCV*.

[11] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *IVC, Special Issue on Human Behavior*, 2009.

[12] A. Gijsenij, T. Gevers, and J. van de Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 2011.

[13] I. Jolliffe. *Principal Component Analysis*. 1986.

[14] M. Knapp and J. Hall. *Nonverbal Communication in Human Interaction*. 2009.

[15] Q. Liu and E. Prakash. The parameterization of joint rotation with the unit quaternion. In *DICTA*, 2003.

[16] M. Livne et al. Human attributes from 3d pose tracking. *CVIU*, 2012.

[17] A. López-Mendez, M. Alcoverro, M. Pardàs, and J. R. Casas. Real-time upper body tracking with online initialization using a range sensor. In *ICCV Workshops*, 2011.

[18] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

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

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

[21] S. Mitra and T. Acharya. Gesture recognition: A survey. *SMC*, 2007.

[22] L.-P. Morency, I. K. de, and J. Gratch. Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *ICMI*, 2008.

[23] P. Natarajan, V. K. Singh, and R. Nevatia. Learning 3d action models from a few 2d videos for view invariant action recognition. In *CVPR*, 2010.

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

[25] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[26] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 2011. Published online December 2011.

[27] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011.

[28] C. Scheffler and J.-M. Odobez. Joint adaptive colour modelling and skin, hair and clothes segmentation using coherent probabilistic index maps. In *BMVC*, 2011.

[29] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[30] L. Sigal and M. Black. Guest editorial: State of the art in image- and video-based human pose and motion estimation. *IJCV*, 2010.

[31] R. Sotil. *Motion models for robust 3D human body tracking*. PhD thesis, 2006.

[32] P. A. Viola and M. J. Jones. Robust real-time face detection. In *ICCV*, 2001.

[33] A. Yao, J. Gall, L. V. Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1359–1367. 2011.

[34] L. Yebin et al. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, 2011.