Let Your Body Speak: Communicative Cue Extraction on Natural Interaction using RGBD Data

Alvaro Marcos-Ramiro, *Member, IEEE,* Daniel Pizarro, *Member, IEEE,* Marta Marron-Romera, *Member, IEEE,* and Daniel Gatica-Perez, *Member, IEEE*

***Abstract*—Employment interviews are relevant scenarios for the study of social interaction. In this setting, social skills play an important role, even though the interactions between potential employers and candidates are often limited. One fundamental aspect of social interaction is the use of nonverbal communication, which affects how we are socially perceived. We present a method to automatically extract body communicative cues from one-on-one conversations recorded with Kinect devices. First, we find the three-dimensional position of hands and head of the subject, and aided by training data, we infer the upper body pose. Then, we use the inferred poses to perform action recognition and build person-specific activity descriptors. We evaluate our system with both domain-specific and public, generic datasets, and show competitive performance.**

***Index Terms*—social interaction, nonverbal cues, markerless motion capture, rgb-d fusion.**

I. Introduction

Social skills play an important role in many aspects of our lives, including the workplace. In job selection processes, employers determine the suitability of candidates through face-to-face interviews, as one of several possible assessment instruments. In job interviews, there is often limited time, both for the candidates to convey their interests and qualities, and for the employers to make judgments. How candidates portray themselves during this short period becomes crucial [1], and defines an interesting subject of study. In particular, nonverbal communication has a strong effect on how we are perceived in social interactions. This matter has been studied in social psychology and cognitive science [2], [3]. In those domains, human coders are used to manually label behaviors and traits. These tasks are typically time-consumming, are subject to inter-coder variations, and face a problem of scalability for big datasets (which could be generated by large firms or assessment centers.)

The main goal of this work is to automatically obtain upper body nonverbal cues that are potentially useful to understand the perception of job candidates in employment interviews. As advantages, the system we propose would bring repeatability to this task, and reduce the time needed to analyze large amounts of data in comparison to human coders.

In order to analyze nonverbal cues, we look for (a) *adaptors*: unconsciously-used movements like nail biting and head scratching, which might provide information about a person's attitude or confidence level; (b) *beat gestures*: movements that do not present a discernible meaning; these are small, rapid flicks of hands and fingers that often beat along with the rhythm of the speech [4], and that can be used to signal points in time when the speaker considers something important relative to a wider discourse [5]; and (c) *posture*: positions of the body (either intentional or habitual) that can be an important clue about the emotional state of people [6], [2].

Several of these cues have been studied in other communicative contexts and are documented at large in the social psychology literature [2]. A few of these cues have also been discussed with respect to the challenges associated with their automatic extraction [7]. Specifically related to the employment interview setting, psychology research has examined the effect of smiling, body posture, and speech cues grouped under the immediacy behavior term, i.e., cues that elicit a perception of closeness and thus a positive impression [8], [9].

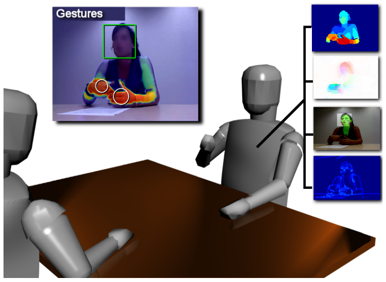


Fig. 1. Using several image and depth cues, our proposed framework outputs hand position, speed, approximate upper body 3D pose and estimated ongoing activity. We use conversational video sequences recorded with Kinect devices as input. Best viewed in color.

Our system uses RGB and depth video sequences showing the upper body of a person during a job interview (see Figure 1). We do not require any body markers, as they could influence the naturalness of the behaviour. This is known as markerless motion capture, and it has been extensively studied before [10]. This problem is hard and challenging when monocular videos are used as input. Recently, range sensors such as Kinect have become accessible to researchers, removing some of the ambiguities and difficulties present in monocular images.

This paper extends the work presented in [11], where the upper-body pose was estimated using monocular video sequences. In this paper we present an upper-body motion capture system based on 2.5D video sequences, which uses multiple image cues, such as face, hand, and motion detectors.

A. Marcos-Ramiro, D. Pizarro-Perez and Marta Marron-Romera are with the Department of Electronics of the University of Alcala, Madrid, Spain. e-mail: {amarcos, pizarro, marta}@depeca.uah.es

D. Gatica-Perez is with Idiap and EPFL, Switzerland. e-mail: gatica@idiap.ch

With this, the system determines 3D upper-body pose and hand speed jointly with conversational actions. Figure 2 shows a general diagram of our proposal. Our method represents a first step towards interaction analysis systems that would help to reduce the burden of human coders.

II. Related work

A. Computational modeling of interaction

Nonverbal behavior [2] has been previously studied video footage. To extract information, these studies require human interpreters with the role of annotating and analyzing data. Automatic nonverbal analysis is an important tool that reduces human annotation in studies that involve a large amount of data. Additionally, automatic annotation is more objective than a human operator. As a result, this problem has attracted the research community and a significant amount of literature has been published on the subject [12], mainly studying interactions in small groups and dyads. Extracting body motion is the main cue in nonverbal behaviour. In wearable computing, there is work that has proposed to attach sensors to the body as a solution to get accurate body motion in human interactions [13]. While motion can be accurate, in these approaches wearing intrusive devices compromises naturalness of the behaviors. Non-invasive sensors, such as cameras, are the preferred approach for researchers. This is however a challenging problem. Basic image features (e.g. visual motion or basic hand gestures) have been studied to obtain human motion, as they can be robustly extracted from video. However, they correspond to rough representations of actual activity [14]. Recent works [15], [16] rely on manual annotations for the body pose in large videos, while automatically extracting other cues such as face expression. This emphasizes the need for automatic body pose detection in video sequences. Hand gesture recognition has also been the subject of a great interest in computer vision [17]. However, these methods are oriented to human-computer interaction or sign language recognition, and therefore their objectives are different to the ones we address here.

Detecting automatically body communicative cues from video sequences falls into a field of computer vision called action recognition. In this work, we recognize specific actions identified as nonverbal cues: *i*) adaptors, which are specific movements, such as head scratching, that provide information about human attitudes and states [4], *ii*) beat gestures, which are flicks of hands used to emphasize important parts of the speech [5] and *iii*) body posture, which can be an indicator of emotional states [2]. In action recognition, the ongoing activity is directly extracted from image cues. Two main approaches can be found in the existing literature. In the first category [18], image cues are usually based on low- and mid-level features such as local space-time features. The activity is then inferred from these features. In the second approach, the body pose is first obtained (i.e. by means of motion-capture), and then used to detect the activity [19], [20]. Even though in the first approach it is possible to perform activity recognition without knowing the body pose, the second approach is generally more accurate [20].

B. Motion capture

Automatic human motion capture is an important problem with a wide range of possible applications. One traditional solution consists on attaching sensors to the body, mainly optical, mechanical or magnetic devices. These approaches are unsuitable in some cases as they constrain body movements influencing their naturalness. More recently, markerless motion capture was proposed, where sensors are placed in the environment, mainly cameras. This problem has been extensively studied [21], [10] and is still considered an open problem. Approaches in markerless motion capture can be divided into three groups: single-camera (monocular) systems, multi-camera systems, and range camera (so-called 2.5D) systems.

Regarding the complexity of the sensor, using a single camera is the simplest option but represents the most challenging problem. In [22], motion capture and action learning is inferred from a few key-pose annotations. In [23], a torso detector is combined with a series of heuristics to infer pose in almost unconstrained still images. In [24], body pose and background segmentation are inferred with high and low level information, together with a coherence term. Our previous work [11] combined hand and head tracking, image motion analysis and incorporated pose priors. This method was specifically designed for upper-body prerecorded sequences.

Multiple camera approaches greatly help to remove pose ambiguities and to obtain the pose in 3D. Although other alternatives exist, the traditional approach when having multiple cameras is to first obtain multi-view silhouettes of the body and then to iteratively adapt a model to these silhouettes [25]. Recently [26] showed that having a rough idea of the action being performed helps motion capture by searching more efficiently through the state space. Other recent approaches include removing motion blur [27], or segmenting and tracking multiple people [28]. The main drawback of multi-view motion capture is that these systems require accurate calibration, camera synchronization and good background segmentation.

Recently, range cameras have been successfully used for markerless motion capture. This is middle case scenario between monocular and multi-camera systems: while only one point of view is available, depth gives relative 3D information. With depth one can obtain 3D poses, reduce ambiguities and gain high invariance to person's appearance. The work of [29], included in the Kinect system, proposed a very accurate and fast pose estimation system. This proposal had enough quality to be used in the videogame industry. It features a body part detector implemented by training a random forest with a huge number of synthetically-generated poses. Some extensions of this work have recently appeared [30], [31], [32], increasing performance and accuracy. A more detailed review of the literature can be found in [33].

The work proposed here is an improvement over the work we first presented in [11], which extracted information from conversational contexts using monocular cameras. In contrast, in this paper a depth sensor is incorporated.

C. Paper contributions and organization

We present next the main contributions of this work:

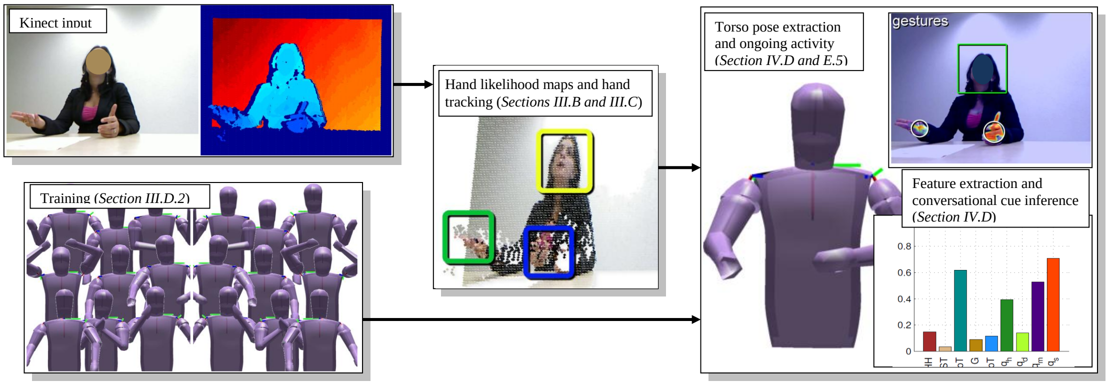

Kinect input
Training (Section III.D.2)
Hand likelihood maps and hand tracking (Sections III.B and III.C)
Torso pose extraction and ongoing activity (Section IV.D and E.5)
gestures
Feature extraction and conversational cue inference (Section IV.D)
0.8
0.6
0.4
0.2
0
HH
ST
oT
G
oT
q_h
q_d
q_m
q_s

Fig. 2. Proposed framework to first extract and then analyze the body posture in conversational sequences. Faces are blurred only for displaying purposes, not for processing.

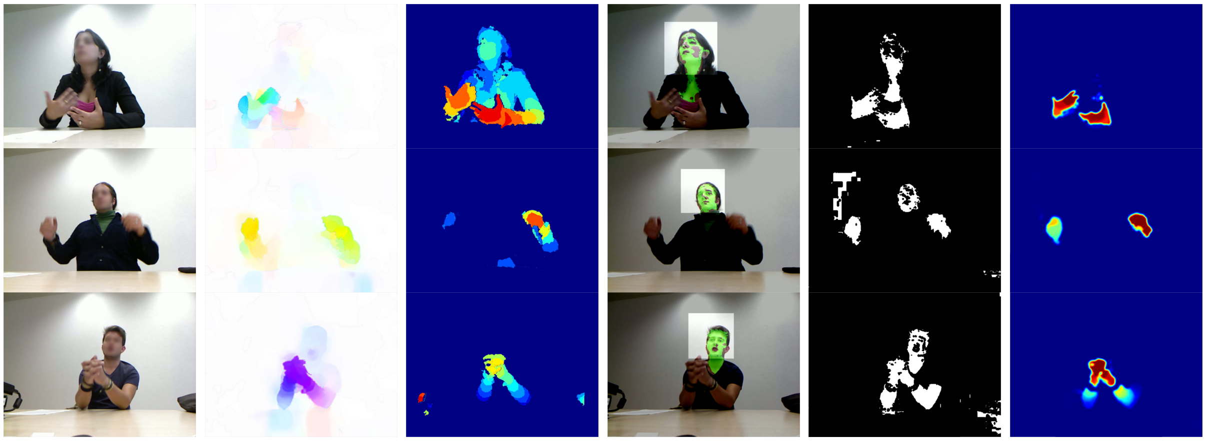


Fig. 3. Steps for building the hand likelihood maps. All images in the same row corresponds to the same time instant. Columns, from left to right: input video frame I, optical flow I_{OF}, depth map I_D, skin segmentation and face ROI (I_S, I'_F), and hand likelihood map I_H (the intersection of the other cues). Best viewed in color.

1) We present an RGBD-based method for hand tracking that notably improves [11]. We detect hands as the body parts that are closest to the camera. This information is available thanks to the depth image. We also improve the analysis of the hand likelihood map for better hand position extraction.

2) We improve the performance of [11] in terms of pose and action retrieval. We use non-linear optimization and we improve robustness of our method for action recognition.

3) We improve our method's evaluation with a dataset in which 27 real job interviews were recorded, accounting for over 4.5 hours.

The rest of the paper is organized as follows: in section III, we present the proposed method for hand mapping, hand tracking, and torso pose estimation. In section IV, we evaluate these algorithms using several databases, and discuss the limitations of our framework. Finally, in section V we present conclusions.

III. Proposed method

A. Overview

Given a RGBD sequence of a person's upper body, we propose several steps for nonverbal cue analysis. First, by combining RGB and depth information, we build a hand likelihood map from which we get the position of both hands along the sequence. Second, we use a head detector and training data to obtain the three-dimensional pose of the torso. Finally, we obtain a person-specific descriptor that models the basic nonverbal behavior of the person in the sequence. See Figure 2.

B. Hand likelihood maps

Given an uncalibrated RGB video sequence $\mathbf{I}_t$ and its corresponding range image $\mathbf{I}_{D,t}$ at time indices $t = t_0 ... t_f$, we obtain the position of both hands in the video. We use a combination of several image cues, which should be as color/appearance invariant as possible to increase robustness. They should also take advantage of the specific constraints of a face-to-face upper body setting.

Our hypothesis is that (even if not necessarily true for every instant) while taking into account a whole sequence, hands are the parts of the image that show most motion and are closest to the camera, in the context of a controlled enviroment in which the subjects are directly facing the camera. Because we take the dynamics of the sequence into account, it models a degree of robustness against short movements that violate these assumptions, such as pose adjustements wearing short-sleeve T-Shirts. Two strong indicators are, respectively: *a*) hands are the furthest body part from the body's axis of rotation, so they show the highest spatial speed for a given joint angular speed, and *b*) the nature of this specific setting with a frontal point of view shows a tendency of orienting the arms and hands closer to the camera than the rest of the body.

In order to formalize these hypotheses, we built a hand likelihood map, where numerical values are proportional to the expectancy of a hand being in that region. The hand likelihood map follows the assumption that, in an image, the hands are skin-colored parts, show more amount of motion, and are closer to the camera. In order to enforce this, we need to compute the optical flow of the sequence to extract motion information, skin segmentation, and face detection. Also, given the natural appearance of the fingers, which have lots of edges, we explored image edge detection as a feature. We detail the steps in Figure 3 and in the rest of this section. In parallel to this work and after our first submission, [34] presented a hand saliency map which uses a very similar concept to ours.

1) Image motion retrieval: We use a state of the art optical flow estimation framework [35] to retrieve image motion. It provides smooth optical flow (see Figure 3, second column) by performing convex optimizations while being resistant to outliers. We use the optical flow modulus $\mathbf{I}_{OF,\rho,t}$ as an input to our method.

2) Face detection: In order to detect the face of the conversing person in the video, we employed a probabilistic version of the Viola & Jones face detector [36]. This method uses likelihood information from the output of every Adaboost cascade classifier, so that the output is probabilistic rather than binary. An initial mask $\mathbf{I}_{F,t}$ is set to 0 inside the face region of interest and 1 otherwise. With this mask we can discard pixels belonging to the face when computing the hand likelihood map, as they are also skin-colored. This approach may not be enough with subjects wearing open neck clothes, resulting in skin-colored pixels falling outside the face detector bounding box, which could be considered as belonging to hands. To address the problem, we take several skin-colored pixels within the face bounding box, and use them as seed points for a region-growing segmentation algorithm, which employs an RGB similarity criterion. The growth is stopped in depth discontinuity points in order to account for the possible inclusion of hands in the growing region, as they are on a different depth level than the face. The output is a binary image $\mathbf{I}'_{F,t}$. As seen in Figure 3, 4th column, this method gives very accurate face and neck segmentations.

3) Edge detection: We use a simple Canny edge detector 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 real-valued edge map $\mathbf{I}_{E,t}$.

4) Skin segmentation: Inspired by [37], we use face detection to infer skin color, as the hue values of face and hands are usually similar. After having processed the face detections in the sequence, a number of $n_{f,S}$ frames frames are chosen randomly from it. Then we analyze the face ROI.

As established in [38], skin color hue values usually fall within the $(0,0.2)$ range of the hue channel in an HSV image. Therefore we set all hue values that satisfy that constraint to be skin candidate pixels. Using this notion, we get the color statistics of the candidate pixels, in order to get their mean and standard deviation of their hue (μ_S, σ_S), which constitutes our subject-specific skin model. After this has been computed, a per-pixel Mahalanobis distance is computed and thresholded for every input image $\mathbf{I}_t$ to get a binary segmentation, which is later refined with simple morphological operations. The result is the binarized result image $\mathbf{I}_{S,t}$.

5) Depth: We retrieve real-valued range images $\mathbf{I}_{D,t}$ by using a commercial Microsoft Kinect sensor. The background is segmented with a distance threshold, while the table is originally undetected because of the high angle of attack relative to the infrared beam of the range sensor. The resulting segmented torso can be seen in Figure 7. It should be noted that we inverted the range values so that closer parts relative to the camera have higher numerical values. Depth and RGB images are registered, therefore their pixels refer to the same points.

6) Hand likelihood map formation: The hand likelihood map is obtained as a combination of these cues. We study several merging strategies, to account for the assumptions explained in Section I.

1) Baseline configuration, comparable to that of [11].

$$\mathbf{I}_{H,t} = \mathbf{I}_{OF,\rho,t}\mathbf{I}_{S,t}\mathbf{I}_{F,t}\mathbf{I}_{E,t}. \tag{1}$$

2) Combined depth and optical flow, removed edges, improved hand detection.

$$\mathbf{I}_{H,t} = \mathbf{I}_{S,t}\mathbf{I}'_{F,t}(\mathbf{I}_{OF,\rho,t} + \mathbf{I}_{D,t}). \tag{2}$$

3) Added depth, weighted optical flow and depth fusion, and new face region segmentation, with removed edges.

$$\mathbf{I}_{H,t} = \mathbf{I}_{S,t}\mathbf{I}'_{F,t}(\kappa_1\mathbf{I}_{OF,\rho,t} + \kappa_2\mathbf{I}_{D,t}), \tag{3}$$

where the constants κ_1 and κ_2 are empirically chosen and used to balance the importance of the optical flow and depth. See Figure 3 for an illustration of equation (3).

C. Hand tracking

In this paper we work with prerecorded video sequences. This is a reasonable assumption when analyzing job interviews (see Section I). We exploit that to track the hands taking into account the whole sequence. The required steps are shown in Figure 5.

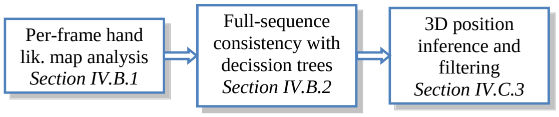
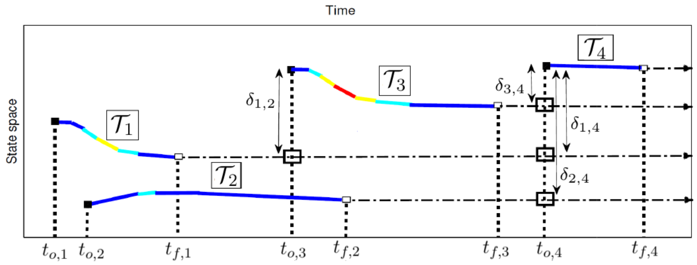


Fig. 5. Hand tracking framework.

1) Finding local maxima in $\mathbf{I}_{H,t}$: For each hand likelihood map frame $\mathbf{I}_{H,t}$, we perform a search for local maxima. We first regularize the likelihood map with a smoothing filter, obtaining $\mathbf{I}'_{H,t}$. We then use Mean Shift to find the different modes of $\mathbf{I}'_{H,t}$ (See Figure 6). The number of modes is not restricted in this step. We also provide identity consistency for the several local clusters along time. At this point we have a set of local maxima of the whole sequence of hand likelihood maps.

Aided by the identity consistency, we compute the paths of the modes in $\mathbf{I}'_{H,t}$ over time. This originates a set of $n_{\mathscr{T}}$ trajectories through the hand likelihood map. We call them tracklets, and are non continuous in the sense that detected



$$\lambda_{T,1} = \lambda_1$$
$$\lambda_{T,2} = \lambda_2$$
$$\lambda_{T,3} = \lambda_3$$
$$\lambda_{T,4} = \lambda_4$$
$$\lambda_{T,1,3} = \lambda_1 + \lambda_3 e^{-\kappa_d \delta_{1,3}}$$
$$\lambda_{T,1,4} = \lambda_1 + \lambda_4 e^{-\kappa_d \delta_{1,4}}$$
$$\lambda_{T,2,4} = \lambda_2 + \lambda_4 e^{-\kappa_d \delta_{2,4}}$$
$$\lambda_{T,3,4} = \lambda_3 + \lambda_4 e^{-\kappa_d \delta_{3,4}}$$
$$\lambda_{T,1,3,4} = \lambda_1 + \lambda_3 e^{-\kappa_d \delta_{1,3}} + \lambda_4 e^{-\kappa_d \delta_{1,4}}$$

Fig. 4. Left: Hand tracking tracklet decision tree example with 4 tracklets ($\mathscr{T}_1$ - $\mathscr{T}_4$) 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.

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 group of tracklets is defined as follows:

$$\mathscr{T} = \{\vec{t_i}\}_{i=1}^{n_{\mathscr{T}}} = \{[t_{o,i}, t_{f,i}, \lambda_i, \vec{u_{o,i}}, \vec{u_{f,i}}]\}_{i=1}^{n_{\mathscr{T}}}, \tag{4}$$

where $n_{\mathscr{T}}$ is the number of tracklets in the sequence; $[t_{o,i}, t_{f,i}]$ are the time instants when the tracklet i starts and ends; λ_i is the accumulated likelihood along the tracklet i duration, $\vec{u_{o,i}}$ is the pixel position where the tracklet i started, and $\vec{u_{f,i}}$ is the pixel position where the tracklet i ended. Longer tracklets therefore usually have bigger λ_i. As tracklets do not have a maximum length value, λ_i is not upper-bound.

2) Full-sequence consistency with Decision Trees: 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 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. This is key to enforce the assumption that $\mathbf{I}_{H,t}$ encodes the most likely hand trajectories when taking the whole sequence into account, even though it does not have to hold true for every time instant.

- 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 to 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 $\vec{u_{o,i}}$ of $\mathscr{T}_i$ to the final point $\vec{u_{f,j}}$ of $\mathscr{T}_j$ through points $\vec{u_{f,i}}$ and $\vec{u_{o,j}}$), separated by a distance $\delta_{ij} = \|dist(\vec{u_{f,i}}, \vec{u_{o,j}})\|$, is:

$$\lambda_{T,ij} = \lambda_i + \lambda_j e^{-\kappa_d \delta_{ij}}, \tag{5}$$

where κ_d is a distance penalization factor (manually set in experiments). We then look for the path with the highest accumulated likelihood. As the used job interview sequences are long (some up to more than 20 minutes), the number of tracklets can be large.

Tracking two hands: As there are two hands to track, we look for the two trajectories (i.e. tracklet paths) with the highest likelihoods. We first define a priority hand, that is, the one that will evaluate the tracklet tree first, thus getting the best path. After it 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 (left or L and right or R) in the image at time t:

$$\mathscr{H}_{L,R} = \{\vec{h_{L,R}}\}_{i=1}^{n_f} = \{[\vec{u_{L,R,i}}, t_i, \lambda_i]\}_{i=1}^{n_f}, \tag{6}$$

where n_f is the total number of frames of the sequence.

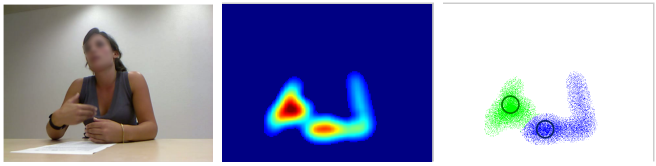


Fig. 6. Hand likelihood map local maxima search with Mean Shift. Left: Original image. Middle: Regularized hand likelihood map. Right: Modes found with Mean Shift (in different colors) and their centroids, represented with circles. Best viewed in color.

3) 3D position inference and filtering: Up to this point, the 2D hand paths are stored in the tracklets, while the 2D head position is obtained with a Viola & Jones detector, as explained in Sections III-B and III-C. We use the depth value of the 2D tracks, obtained with the range sensor, in order to infer the hands and head 3D positions. As Figure 7 shows, regions of interest around the 2D locations are analyzed. Hands are deemed to be the closest point to the camera within their respective regions, and the head the furthest point. This is justified as after segmenting $\mathbf{I}_{D,t}$ the wall and table are excluded, so there is nothing behind the head or in front of a hand.

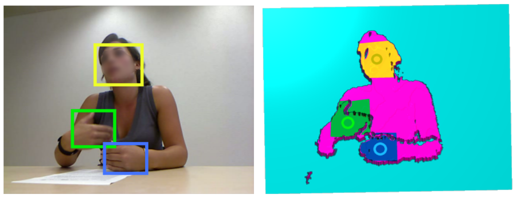


Fig. 7. Segmented depth with its hand and head location. Left: Bounding boxes of the 2D hands and head locations. Right: Search for the 3D points in the depth image. Best viewed in color.

The information stored in the tracklets (both hand position and number of hands detected) can be noisy due to the tracking-by-detection nature of the system. In order to address this, we encode the number of detected hands as states in a Hidden Markov Model. By using the Viterbi framework, and establishing a tendency to stay in the current state (90% in the transition probability matrix) we get a more stable hand count. Finally, we implemented a Kalman Filter to get the final 3D hand path.

D. Torso pose extraction

In order to infer the torso 3D pose, we propose to adjust an articulated upper body model to the head and hands 3D positions. Our method requires training data. To build this data, we first collect and label several typical conversationally-relevant upper body poses with a range camera. Then we use a non-linear optimization technique to minimize a joint energy function where trainning data is used to constrain the solution. The process is explained as follows, and summarized in Figure 8.

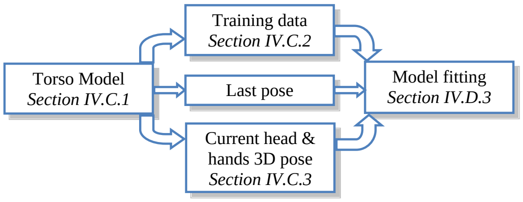


Fig. 8. Torso pose extraction overview.

1) Torso model: We use a synthetic 3D polygonal torso mesh model, driven by an underlying skeleton with n_{joints} joints (see Figure 9). The skeleton pose $\mathscr{A}$ is parameterized by the 3D euclidean rotation angles:

$$\mathscr{A} = \{\vec{a_i}\}_{i=1}^{n_{joints}} = \{a_{\alpha i}, a_{\beta i}, a_{\gamma i}\}_{i=1}^{n_{joints}}. \tag{7}$$

The angles $[a_{\alpha i}, a_{\beta i}, a_{\gamma i}]$ correspond to pitch, yaw and roll, and are applied in a hierarchical manner. That is, to obtain the orientation of a given body part, angles must be composed in chain relative to the root node (the base of the neck joint). The root node is referenced to the world global coordinates by its 3D position and orientation.

We have not experienced any problems regarding Gimbal locks, therefore we did not consider necessary to change to other rotation representations, such as quaternions.

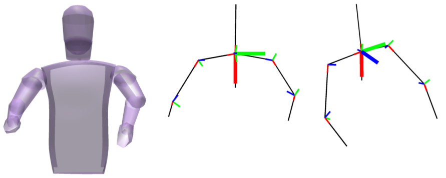


Fig. 9. Torso Model. Left: 3D mesh. Center and right: underlying skeleton model. The base of the neck is the root node. Red, green and blue lines correspond to the X, Y and Z axes, respectively. Best viewed in color.

2) Training data: Just using the hands and head positions to fit the torso can lead to unrealistic poses. Therefore, we take advantage of prior information gathered by an offline training process to constrain the possible space of poses. This offline process has to be performed only once. Four subjects (two male, two female) are recorded with a range camera in a similar setting to the target scenario (i.e. seated at the table, see Figure 10.A left), while performing a set of N_{Ta} actions, resulting in a total of N_T training frames. As seen in Figure 10.C, the actions are all typical of conversational settings. The objective is to capture the adaptors, beat gestures and body pose information.

The actions are recorded with a range camera, retrieved a set of $\mathbf{I}_{D,tr}$ training range images. These images are then manually annotated to label the joint position. For example, an annotation for joint i is expressed as:

$$\vec{p_{i,tr}} = [\vec{u}, \mathbf{I}_{D,tr}(\vec{u})] = [u, v, \mathbf{I}_{D,tr}(\vec{u})], \tag{8}$$

and is therefore described as its pixel position in the image and its associated distance from the camera, thus forming a 3D vector. Manually annotating the position of every joint along all the training depth recordings allows us to obtain the relative 3D location of the whole body, forming the set of training poses:

$$\mathscr{P}_{tr} = \{\vec{p_{i,tr}}\}_{i=1}^{n_{joints}} = \{x_{i,tr}, y_{i,tr}, z_{i,tr}\}_{i=1}^{n_{joints}}. \tag{9}$$

If an occlusion occurs, we set the position of the occluded joint either as the one used in the last frame, or as an estimated guess. As seen in Figure 10.B, we add joint angle energy functions constrains for imposing pose naturalness. Even if rough, this setting produces good results for obtaining the desired parameterization (see Figure 10.A, and 3D mesh in Figure 9).

3) Model fitting: In order to infer the torso 3D pose, we propose to adjust an articulated upper body model to the head and hands 3D measurements, helped by training data. As explained before, to build this data we first collect and label several typical, conversationally-relevant, upper body poses with a range camera. Then we use a non-linear optimization technique to minimize a joint energy function. The process is explained as follows, and summarized in Figure 8.

The 3D torso pose is extracted in a two-step process. First, an approximate pose is quickly estimated via database lookup, getting the best match by using 2D hands and head position and silhouette as cues. The silhouette has been segmented by identifying the depth blob corresponding to the detected face of the subject. Then, this first guess is used to initialize a nonlinear least-squares optimization method that further refines the pose (see Figure 16). The cost function used is:

$$\varepsilon \propto \|\vec{p}_t - \vec{p_{o,t}}\| + \|e_{n,t}\| + \|\mathscr{P}_{o,t} - \mathscr{P}_{o,t-1}\|, \tag{10}$$

where the term $\|\vec{p}_t - \vec{p_{o,t}}\|$ comprises the 3D hands and head position difference between those detected in the image and the current position in the articulated model. The term $\|e_{n,t}\|$ penalizes the least natural positions by using the function described in Figure 10.B. The term $\|\mathscr{P}_{o,t} - \mathscr{P}_{o,t-1}\|$ is the

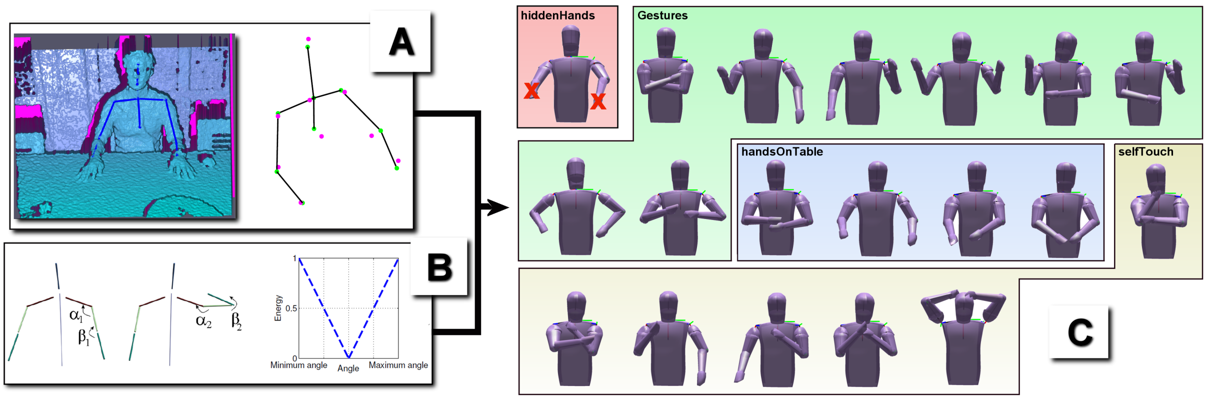


Fig. 10. Training process. A: Labeling. Left: manually annotated 3D skeleton overlaid into the range camera 3D measurements. Right: optimized torso pose relative to the manually annotated points (in magenta). B: Natural pose attainment. Left skeleton: low energy arm pose. Right skeleton: high energy 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. C: Set of trained movements and their classification into four categories. Best viewed in color.

difference between the current estimate and that of the previous frame. This term is used for temporal consistency. After this process, the approximate 3D upper body position of the participant $\mathscr{P}_o$ is retrieved.

E. Feature extraction and conversational cue inference

At this point we have obtained the hands' position in the image and the approximate 3D torso pose. As the next step, we extract a series of features from the participant, that can schematically be seen in Figure 11.

1) Hand height ($q_{h,t}$): By using a Hough detector, we obtain the table edge position in the image. We then compute the distance in pixels between the edge and the face of the participant. The height of each hand is expressed as a proportion relative to the face-table distance. A value of 0 means the hand is located in the edge of the table, while a value of 1 means that the hand is at the same height of the face. If the hand has not been detected, a -1 value is assigned. This feature is therefore two-dimensional (1D per hand).

2) Hand movement ($q_{m,t}$): The detected hand position time differences are not a reliable indicator for hand speed, since the tracker can focus on different parts of the hand, leading to inaccuracies. Also, it does not capture the nuances of small hand and finger movements, which are still hand activity. In order to circumvent that problem, we used the average optical flow modulus present in a region of interest around the detected hand coordinates as hand speed measure. This feature is also two-dimensional.

3) 3D face-hand distance ($q_{d,t}$): We compute the euclidean 3D distance between face and hands, and normalize it with respect to the face-table distance. Lower values indicate closeness to the face. This feature is two-dimensional.

4) Speaking status ($q_{s,t}$): The commercial microphone array Microcone[1] provides automatic binary speaking status segmentation (talking or silent), that we use as a feature.

5) Ongoing activity ($q_{a,t}$): Five classes were defined based on the occurrences in the dataset and the relevance in the nonverbal communication literature [2], as Table I shows: "Hidden Hands", "Hands on table", "Gestures on table", "Gestures", and "Self-touch". Self-touch is a nonverbal cue largely studied in nonverbal communication [2] and in multiple interpersonal situations. Some self-touch cues are sometimes called self-intimacies, and include touching one's face, "holding one's own hands, arm folding, and leg crossing" ([2], p. 253.) In our case, we are interested in detecting face self-touch. Self-touch cues have sometimes been linked in the literature with "situational anxiety or stress" ([2], p. 255) and so they are relevant to detect in the interview setting. Gesture and Gesture on table, as we define them, are accounting for a number of classes that span adaptors and beat gestures. Finally, Hands on table and Hidden hands are clearly context-dependent categories, and valid only for interactions where such physical artifact is present, which creates a framework in terms of interpersonal distance that imposes constraints in the expression of proxemics [2].

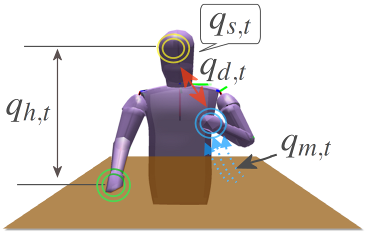


Fig. 11. Graphical representation of the per-frame extracted features: hands height ($q_{h,t}$), hands movement ($q_{m,t}$), 3D face-hands distance ($q_{d,t}$) and speaking status ($q_{s,t}$). Best viewed in color.

These cues constitute an approximation for the applicants' body posture and gestures, as with applicants seated, the posture was for a large part defined by the position of their arms. Other posture classes such as leaning forward or backward were also considered, but were discarded as the observed variability of such postures was low.

Up to this point we have therefore a 7-dimensional feature vector for instant t:

$$\mathscr{Q}_t = \{q_{h,t}, q_{m,t}, q_{d,t}, q_{s,t}\}. \tag{11}$$

[1] http://www.dev-audio.com/products/microcone/

In order to infer the ongoing activity ($q_{a,t}$), we train a Random Forest classifier by concatenating $t+t_{off}$ and $t-t_{off}$ feature vectors $\mathscr{Q}$, where t_{off} denotes a time offset. We then obtain the final $(7 \cdot 2 \cdot t_{off}+1)$-dimensional feature vector for time instant t, that we use for training:

$$\mathscr{Q}'_t = \{\mathscr{Q}_{t-t_{off}:t+t_{off}}\} \tag{12}$$

The labels for training are a set of manual annotations of the perceived ongoing activity along the whole video corpus (see Section IV.A for details) The output is therefore a $q_{a,t_0:t_f}$ vector that encodes the action performed in every time instant t. Its performance is evaluated in Section IV.C.

6) Final nonverbal cue ($\mathscr{X}$): Combining all the previous features, we obtain a summary of the ongoing activity of the participant for the interaction:

$$\mathscr{X} = \{mn(q_{h,t}), mn(q_{m,t}), mn(q_{d,t}), mn(q_{s,t}), mn(q_{a,t})\}, \tag{13}$$

where mn is the mean function. The resultant feature $\mathscr{X}$ is nine-dimensional, since we also average the hand-related features for both hands. Therefore $mn(q_{h,t})$, $mn(q_{m,t})$ and $mn(q_{d,t})$ become one-dimensional. The feature $mn(q_{a,t})$ is five-dimensional since it comprises an action frequency histogram for each of the categories listed in Table I.

TABLE I
ACTION CLASS DESCRIPTIONS AND RELATIVE FREQUENCY.

Class	Description	Freq.
hiddenHands (HH)	No hands visible in the image	4.97 %
selfTouch (ST)	Touches of face, hair or torso with one or both hands.	11.75 %
handsOnTable (HoT)	Resting the hands in the table	51.89 %
gestures (G)	Gesturing while the hands are not close to the table	11.56 %
gesturesOnTable (GoT)	Gesturing with the hands while they are close to the table, or the arms resting on it	19.83 %

These category definitions are created according to the mentioned literature and after discussing with psychologists what would be the most meaningful action units to detect in the application of interest, and relative to their frequency of appearance in videos. Their ultimate goal is to serve as a proxy to adaptors (which provide information about the psychological state of the sender), beat gestures (which can show what the speaker considers important) and posture (which encodes information about the emotional state), as discussed in the introduction.

IV. Implementation and Results

A. Implementation and data

There is a lack of public databases for testing activity recognition and hand tracking in a seated, conversational, long sequence setting. Therefore, we built a set of experiments in order to test the performance of the hand tracking and action recognition algorithms.

We use data obtained from real job interviews, recorded in a psychology university conversation room (see Figure 12) using two uncalibrated but synchronized commercial Kinect sensors: one pointing at the interviewer and another one at the person being interviewed. They are fixed on a table and pointing at the upper body of the participants (see Figure 1). Images have a resolution of 640 × 480 pixels and video runs at 30 frames per second. The audio is also recorded, by using a Microcone commercial device. For the present work. we focus on the behavior of the person being interviewed, resulting in a total of 27 interviews (7 male and 20 female subjects) and almost 4.5 hours of audio-visual data. The average interview length is 9.8 minutes. The behavior of the person being interviewed is completely natural, and the clothing has not being constrained in any way (see Figure 17).

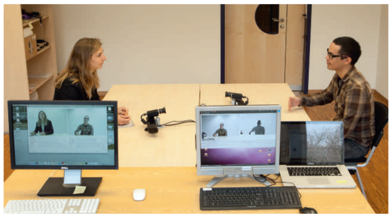


Fig. 12. Video recording setup.

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. That is, we do not process the frames which do not show enough change. However we take them into account when computing the average features. This resulted in 1 hour and 56 minutes of video. Class distribution is shown in Table I. The majority class is 'hands on table' with more than half of the data. The least frequent class is 'hidden hands'.

B. Experiment definition

We evaluate the proposed framework along different parts of the processing pipeline. We compare the results to those obtained in [11], the most similar work to the proposed here. The experiments consist of:

Experiment 1: Hand likelihood map quality evaluation. We evaluate how well the hands are detected by using different hand likelihood map configurations. We use Mean Shift to obtain the local maxima of the likelihood map and we identify them with the hands. Then, we evaluate the detection error with 2 minute sequences (3750 frames and 4 different subjects), in which the position of the hands in the images was manually annotated at every frame. Inspired by [32], we define a detection rate metric in which a hand is considered as 'detected' if a local maximum of the hand likelihood map is found within a given pixel threshold of the labeled position. Different thresholds are explored, and by visual inspection we have found 40 pixels to be the boundary of an acceptable detection. We also analyze the false positive rate (how often a hand is detected in a part of the image in which it is not present) and the missed positive rate (how often a hand detection is missed).

Experiment 2: Generalization of performance. We use the public database ChaLearn 2011, which is non-conversational, but allows to validate our proposal. It contains 437 non-time-consecutive, 320×240 color and depth frames, in which body joints have been manually annotated. The environment is uncontrolled with different backgrounds, high variance of poses, clothing, positioning, and lighting. In some of the frames, there is no movement at all. The metrics that we use to evaluate the performance are *a*) detection rate as defined in Experiment 1, and *b*) maxima average order. In *b*), the local maxima of the hand likelihood map are obtained as described in Section III-C1, and ordered with respect to their likelihood value. Then, each hand annotation is associated with the closest (in its 2D position) maximum. For example, if the right hand is associated with a local maximum which contains the 3rd highest likelihood, and the left hand is associated with the best local maximum, the maximum average order is 2 for that frame. A maximum average order of 1.5 provides the best possible scenario, as the two hands would be associated with the two maxima that contain the highest likelihoods.

Experiment 3: Action recognition performance. In order to evaluate the action recognition algorithm, we manually labeled the actions performed by the 27 different subjects, according to the categories in Table I. To simplify the process, we labeled one every 15 frames (or half a second) in the portions of the video which showed enough motion. This resulted into 13900 manually labeled frames, see Table I for how they split between the different actions. In order to assess the reliability of annotations, a second person annotated 63 minutes of the dataset (around 5000 frames), resulting in a satisfactory interrater agreement of Cohen's Kappa 0.81. As performance measure, we use frame classification accuracy.

C. Results and discussion

Hand likelihood map quality evaluation: The results of the hand detection for *Experiment 1* are shown in Figure 13. We set 40 pixels as a threshold value, which visually is the highest drift accepted for a perceived correct detection. Our method (RGBD+RG) is tested both with and without Decision Trees (DT) tracking. Our method is compared with our previous work [11], and also with the result obtained when leaving out the face skin region growing (RG) procedure (see Section III.B2, and Section III.B.6 for details). It is clear that the addition of depth helps the hand detection in a high degree. Our method gets an average **28.7%** higher detection rate than [11]. The inclusion of RG increases the detection rate in an average of **5.6%**. The overall RGBD+RG detection rate is **78.2%** when using DT. We compare our method with a state-of-the-art detector [39] using our dataset. From Figure 13 we see that detection results are comparable, but at the expense of a much higher false positive rate.

Upon visual inspection, most of the non-detections are a cause of the person's hands being close together, being therefore detected as a single one. In practice, however, this is not a serious issue, as for action recognition usually the ongoing action is a function of where the highest hand is (for example, when gesturing or self-touching, it is irrelevant to leave a hand in the table for the action to be identified). Also, for upper body pose, the regularization term in the optimization method (see Section IV.D.3) usually dampens misdetections. False positives on the other hand are quite costly. If a hand is mis-detected in the face region for several consecutive frames, there is high probability for the action to be incorrectly classified as 'self touch'. This is where the RG algorithm comes into effect. As seen in Figure 13.B, it improves the false positive rate by **16.9%**. The same effect appears when applying the DT tracking scheme: while the outcome of the precision is largely unaffected, it reduces false positives by an average of **12.8%**. It therefore becomes a trade-off between missed hand rate and false positive rate. As the next steps of the pipeline are more affected by false positives, the RGBD+RG with DT becomes the best overall performer: while it misses hands more frequently than RGBD alone, it keeps the more relevant false positive rate much lower while offering better precision.

The results of the hand detection for *Experiment 2* are shown in Figure 14. We use [40], [41], as baselines, together with the body part classification of [29] coupled to a 2D hand position regressor. As expected, methods that use entirely or partially depth outperform the 2D-only methods. In addition, our method is able to locate the hands very precisely, as the comparison with [29] shows. The maximum average order is **2.4** for the right hand and **2.55** for the left hand, showing that in general, one of the best 3 maxima of the hand likelihood map are overlapped with the hand's positions. This enables high quality information to be passed to the tree-based tracker in order to reliably obtain the hand's position along time, as assessed in *Experiment 1*.

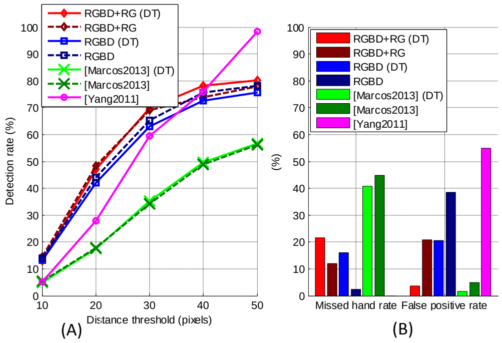


Fig. 13. Hand detection performance. (A) Hand detection rate as a function of the pixel threshold. (B) Missed hand and false positive rates. [Yang2011] corresponds to reference [39].

Action recognition evaluation: The results for action recognition can be seen in Figure 15. Several configurations for the feature vector have been tested. The value of t_{off} has been empirically set to 5, as higher numbers do not provide a significant precision improvement, while increasing the computational cost. All variations of the used feature vector resulted in a significant improvement over the majority class performance. The behavior of the algorithm is consistent along all interviews: under a high variation of clothing, skin color, gender, or class distribution, the mean and standard deviation of the accuracy for the best performing combination

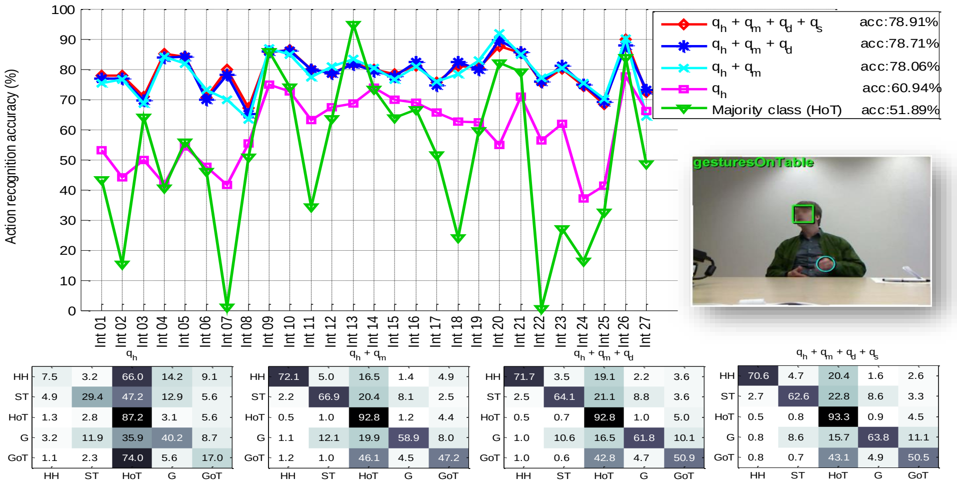


Fig. 15. Action recognition performance. Top: accuracy of the proposed algorithm in the different interviews (Int 1 ... Int 27), in function of different feature vectors configurations, and relative to the majority class performance ('hands on table'). In the right part, a frame of one of the video sequences with the hands, head, and ongoing activity is overlaid. Bottom: confusion matrices for different feature vector configurations. Best viewed in color.

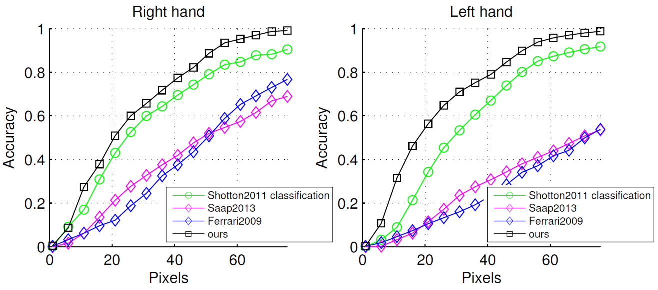


Fig. 14. Results in ChaLearn 2011, as hand detection performance. We compare our method against baselines [40], [41] and a modification of [29].

are **78.9%** and **5.9%**, respectively. The biggest performance jump is found when the motion of the hands $q_{m,t}$ is taken into account: on average, it improved the recognition by **18.9%**. Without this cue, the biggest source of error was the confusion between the two most populated classes, 'hands on table' and 'gestures on table' (see the confusion matrices in Figure 15). Understandably, it shows that hand speed is a big factor to distinguish between those classes. Nevertheless, even when the hand speed is used, the confusion matrices show that mixing both classes is still an issue. However, upon visual inspection of the sequences, the classification is consistent with the amount of movement of the hands. This suggests that the annotators used additional cues other than the amount of movement to distinguish between 'hands on table' and 'gestures on table', such as the orientation of the hand palms or finger position. This possibility offers grounds for future works.

Adding the 3D distance from hands to face ($q_{d,t}$) improved the recognition accuracy, but only by **0.65%**. This shows that although having the 3D position helps, the 2D hands position relative to the face is already enough to distinguish between actions. The same applies for the speaking status ($q_{s,t}$), although it crucially shows that multimodality can be exploited in order to improve action recognition. This finding is also supported with other recent works like [42].

Overall, the present work improves our previous work [11], with a global accuracy and majority class of **78.9%** and **51.9%**, in contrast to the **72.5%** and **67.5%** of [11], while adding an extra 'gestures on table' class. It is important to mention that correctly classifying the majority class 'hands on table' action is not trivial, as factors like slow motions, skin colored clothes, or sleeve-less shirts have to be dealt with. As an illustration, we show some failure examples of the hand tracking in Figure 17. Note that in some cases, even if the hands are wrongly detected, the temporal features used for action recognition recover the right action.

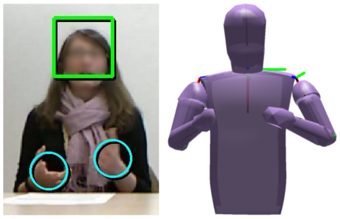


Fig. 16. 3D upper body retrieval. Left: detected face and hands. Right: approximate 3D limbs configuration.

V. Conclusion

We presented a system that automatically analyzes communicative cues of seated participants with commercial RGBD sensors in the context of real job interviews. We built original hand and face detectors to get an approximate 3D upper body pose. With that information, an action recognition system was built by using temporal features. Statistics were then extracted in order to build a feature vector that summarizes key nonverbal behavior present along the whole job interview. Specifically, we look for adaptors and beat gestures, which previous studies have shown to carry nonverbal communication information. Our system can recognize 5 upper-body actions with an accuracy of 78.9%, in a dataset of four and a half hours of real job interviews.

The results obtained with our method have shown to improve over the most related previous work, as tested in one domain-specific database and one public database. Our current aim is to provide more accurate cues to analyze traits of people engaged in conversation. Our future work will continue to investigate the possibilities that these kind of cues provide, and attempt to develop a more fine-grain activity classification framework.

VI. Acknowledgments

This research was funded by University of Alcala FPI program, the SNSF SONVB and UBImpressed projects, and the Spanish Ministry of Economy and Competitiveness under project SPACES-UAH (TIN2013-47630-C2-1-R). We thank Denise Frauendorfer (Université de Lausanne) for support with data collection, and Laurent Nguyen (Idiap Research Institute) for support with data collection and technical discussions.

References

[1] J. Curhan and A. Pentland, "Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first five minutes," *Journal of Applied Psychology*, vol. 92, 2007.

[2] M. Knapp, J. Hall, and T. Horgan, *Nonverbal communication in human interaction*, 8th ed. Cengage Learning, 2014.

[3] A. Pentland, *Honest Signals: How They Shape Our World*, 2008.

[4] D. McNeill, *Hand and Mind*, 1992.

[5] ——, *Gesture and Thought*, 2005.

[6] A. Mehrabian, *Nonverbal Communication*, 1972.

[7] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, "Human computing and machine understanding of human behavior: a survey," in *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 47–71.

[8] A. S. Imada and M. D. Hakel, "Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews." *Journal of Applied Psychology*, vol. 62, no. 3, p. 295, 1977.

[9] R. J. Forbes and P. R. Jackson, "Non-verbal behaviour and the outcome of selection interviews," *Journal of Occupational Psychology*, vol. 53, no. 1, pp. 65–72, 1980.

[10] L. Sigal and M. Black, "Guest editorial: State of the art in image- and video-based human pose and motion estimation," *International Journal on Computer Vision*, vol. 87, no. 1-2, 2010.

[11] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. S. Nguyen, and D. Gatica-Perez, "Body communicative cue extraction for conversational analysis," in *Proc. IEEE Face and Gesture Recognition*, Shanghai, China, April 2013.

[12] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing, Special Issue on Human Behavior*, vol. 27, no. 12, 2009.

[13] S. Feese, B. Arnrich, G. Troster, B. Meyer, and K. Jonas, "Detecting posture mirroring in social interactions with wearable sensors," in *Proc. IEEE International Symposium on Wearable Computers*, San Francisco, CA, USA, June 2011.

[14] 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 of Multimedia*, vol. 14, no. 3, 2011.

[15] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of tv viewers," in *Proc. IEEE Face and Gesture Recognition*, Shanghai, China, April 2013.

[16] S. Scherer, G. Stratou, M. Mahmound, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *Proc. IEEE Face and Gesture Recognition*, Shanghai, China, April 2013.

[17] L.-P. Morency, I. K. de, and J. Gratch, "Context-based recognition during human interactions: Automatic feature selection and encoding dictionary," in *Proc. ACM International Conference on Multimodal Interaction*, Crete, Greece, October 2008.

[18] I. Laptev, "On space-time interest points," *IEEE International Journal on Computer Vision*, vol. 64, no. 2-3, 2005.

[19] J. Gall, A. Yao, and L. J. V. Gool, "2d action recognition serves 3d human pose estimation," in *Proc. European Conference on Computer Vision*, Florence, Italy, October 2012.

[20] G. F. Angela Yao, Juergen Gall and L. V. Gool, "Does human action recognition benefit from pose estimation?" in *Proc. British Machine Vision Conference*, Dundee, August 2011.

[21] R. Urtasun, "Motion models for robust 3d human body tracking," Ph.D. dissertation, 2006.

[22] P. Natarajan, V. K. Singh, and R. Nevatia, "Learning 3d action models from a few 2d videos for view invariant action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010.

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

[24] H. Wang and D. Koller, "Multi-level inference by relaxed dual decomposition for human pose segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Colorado, USA, June 2011.

[25] L. Yebin et al, "Markerless motion capture of interacting characters using multi-view image segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Colorado, USA, June 2011.

[26] A. Yao, J. Gall, and L. Gool, "Coupled action recognition and pose estimation from multiple views," *International Journal on Computer Vision*, vol. 100, no. 1, 2012.

[27] D. Wu, Y. Liu, I. Ihrke, Q. Dai, and C. Theobalt, "Performance capture of high-speed motion using staggered multi-view recording," in *Pacific Graphics*, 2012.

[28] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of interacting characters using multi-view image segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Colorado, USA, June 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Colorado, USA, June 2011.

[30] A. López-Mendez, M. Alcoverro, M. Pardàs, and J. R. Casas, "Real-time upper body tracking with online initialization using a range sensor," in *Proc. IEEE International Conference on Computer Vision Workshops*, Colorado, USA, June 2011.

[31] J. Gall, A. Fossati, and L. J. V. Gool, "Functional categorization of objects using real-time markerless motion capture," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Colorado, USA, June 2011.

[32] J. Taylor, J. Shotton, T. Sharp, and A. W. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, June 2012.

[33] A. Marcos-Ramiro, "Automatic body communication extraction through markerless motion capture," Ph.D. dissertation, 2014.

[34] Y. Yin and R. Davis, "Gesture spotting and recognition using salience detection and concatenated hidden markov models," in *Proc. ACM International Conference on Multimodal Interaction*, Sydney, Australia, 2013.

[35] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, 2011.

[36] F. Sanabria-Macias, E. Maranon-Reyes, P. Soto-Vega, M. Marron-Romera, J. Macias-Guarasa, and D. Pizarro-Perez, "Face likelihood functions for visual tracking in intelligent spaces," in *Proc. IEEE International Conference on Industrial Electronics Society, IECON 2013*, Viena, Austria, November 2013.

[37] C. Scheffler and J.-M. Odobez, "Joint adaptive colour modelling and skin, hair and clothes segmentation using coherent probabilistic index maps," in *Proc. British Machine Vision Conference*, Dundee, August 2011.

[38] A. Gijsenij, T. Gevers, and J. van de Weijer, "Computational color constancy: Survey and experiments," in *Proc. IEEE International Conference on Image Processing*, Brussels, Belgium, September 2011.

[39] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[40] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Pose search: Retrieving people using their pose," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, 2009.

[41] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Ohio, USA, 2013.

[42] L. S. Nguyen, J.-M. Odobez, and D. Gatica-Perez, "Using self-context for multimodal detection of head nods in face-to-face interactions,"

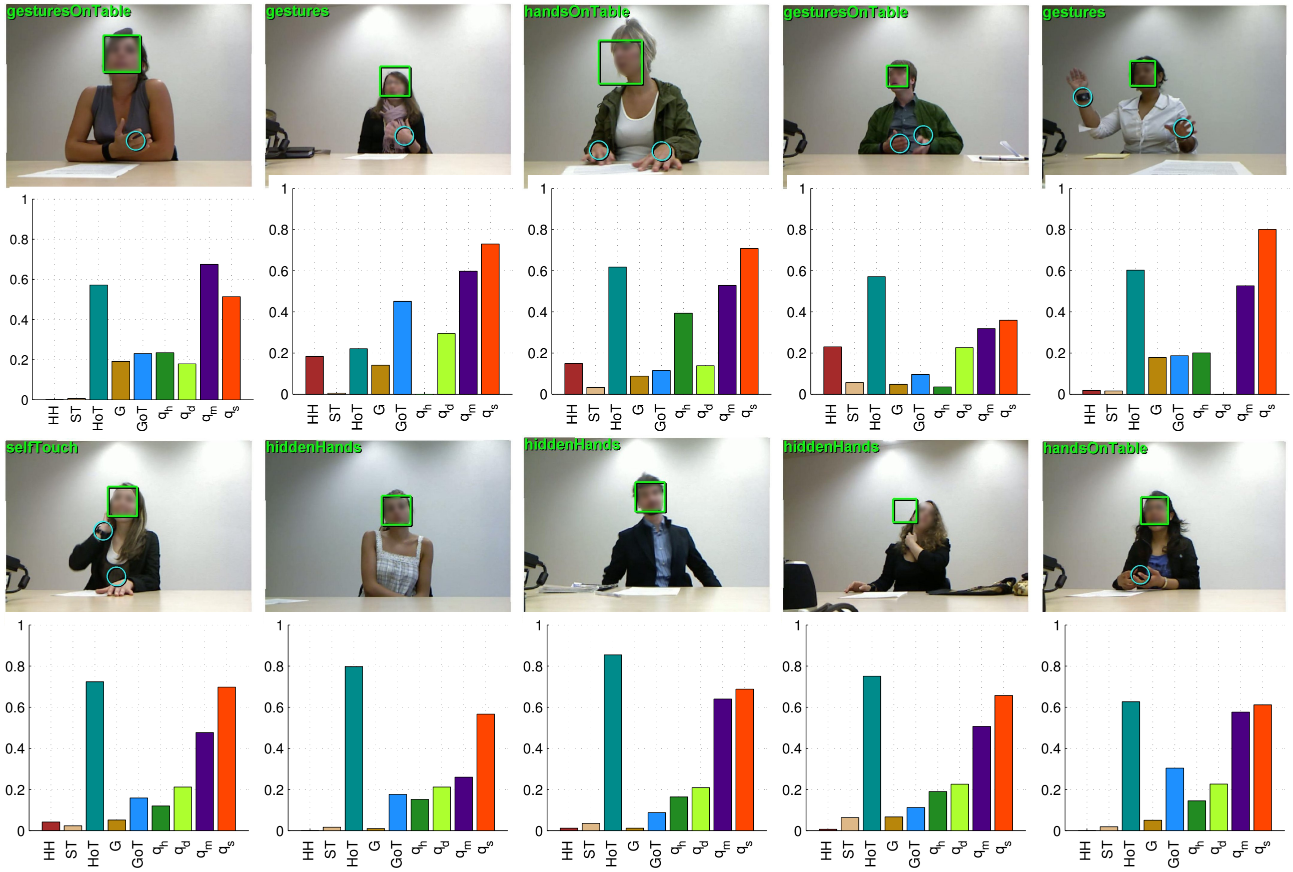


Fig. 17. Frame results, in two rows of frames and corresponding interview descriptor. Overlaid to RGB images are the hand and face position, and the detected ongoing activity. The last two RGB images show failure examples (from left to right: incorrect face detection, merged hands). Best viewed in color.

in *Proc. ACM International Conference on Multimodal Interaction*, California, USA, October 2012.

Alvaro Marcos-Ramiro received the B.Sc. degree in Technical Industrial Engineering, Industrial Electronics in 2007, and the M.Sc. degree in Advanced Electronic Systems in 2009, both in University of Alcala. In 2014, he received the Ph.D. degree in Advanced Electronic Systems from the Electronics Department, University of Alcala, collaborating as a visiting student with the Social Computing Group at Idiap Research Institute, in the fields of markerless motion capture, machine learning, optimization, and nonverbal communication.

Daniel Pizarro received the Ph.D. degree in Electrical Enginneering from the University of Alcala, Spain, 2008. In the period 2005-2012 he was Assistant Professor and member of the GEINTRA group in the Department of Electronics of the University of Alcala. From 2012-2014 he was associate professor and member of ALCoV (Advanced Laparoscopy and Computer Vision) group at the Universite d'Auvergne, France. Since 2014 he is Associate Professor in the University of Alcala. His research topics includes computer vision applied to deformable image registration, detection and reconstruction. He works also for the development of computer vision techniques applied to minimally invasive surgery.

Marta Marron-Romera obtained her degree in Telecommunication Engineering, her masters degree, and her Ph.D. degree in University of Alcala. From 1996 to 2001 she was a granted researcher in the Electronics Department, in the same university, and she was a teacher; since 2009 she has been an Associated Professor. Nowadays she is member of the GEINTRA research group. Her research interests include computer vision, probabilistic algorithms, and robotics. She is an IEEE member from 2004, and member of the RAS, IES and IM IEEE Societies.

Daniel Gatica-Perez is Head of the Social Computing Group at Idiap Research Institute and Professeur Titulaire at the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland. His research interests include social computing, mobile and ubiquitous computing, and social media. He has served as an associate editor of the IEEE Transactions on Multimedia. He is a member of the IEEE.