Capturing Upper Body Motion in Conversation: an Appearance Quasi-Invariant Approach

Alvaro Marcos-Ramiro
University of Alcala
Spain
amarcos@depeca.uah.es

Daniel Pizarro-Perez
University of Clermont-Ferrand
France
dani.pizarro@gmail.com

Marta Marron-Romera
University of Alcala
Spain
marta@depeca.uah.es

Daniel Gatica-Perez
Idiap Research Institute and
EPFL
Switzerland
gatica@idiap.ch

ABSTRACT

We address the problem of body communication retrieval and measuring in seated conversations by means of markerless motion capture. In psychological studies, the use of automatic methods is key to reduce the subjectivity present in manual behavioral coding used to extract these cues. These studies usually involve hundreds of subjects with different clothing, non-acted poses, or different distances to the camera in uncalibrated, RGB-only video. However, range cameras are not yet common in psychology research, especially in existing recordings. Therefore, it becomes highly relevant to develop a fast method that is able to work in these conditions. Given the known relationship between depth and motion estimates, we propose to robustly integrate highly appearance-invariant image motion features in a machine learning approach, complemented with an effective tracking scheme. We evaluate the method's performance with existing databases and a database of upper body poses displayed in job interviews that we make public, showing that in our scenario it is comparable to that of Kinect without using a range camera, and state-of-the-art w.r.t. the HumanEva and ChaLearn 2011 evaluation datasets.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences; G.3 [**Mathematics and Computing**]: Miscellaneous; D.2.8 [**Image Processing and Computer Vision**]: Metrics—*Scene analysis*

General Terms

Monocular motion capture; optical flow; social computing; job interviews

1. INTRODUCTION

Nonverbal communication can influence how we are socially perceived, therefore this subject has been intensively analyzed in social psychology and cognitive science [14]. In these domains, however, there has traditionally been the need for an interpreter. That is, a person that emits a judgment on the perceived traits of the analyzed subject, or that codes specific behaviors from hours of video recordings. This judgment always carries a degree of subjectivity, which can lead to inconsistencies across different evaluations.

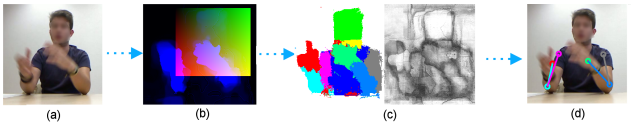


Figure 1: Data flow overview. (a) Original image. (b) Proposed 4-channel image. (c) Body part classification and confidence scores. (d) Obtained pose. Please view in PDF.

Markerless motion capture from monocular images is a good solution to this problem (as markers placed in the body can alter the behavior), but remains a challenge in computer vision. This is a consequence of many factors, such as the high-dimensionality of the data, camera projection distortions, appearance variability (*e.g.* clothing, skin, hair...) or external and self-occlusions. In the last few years range cameras have appeared in the mass market, largely solving the depth perception and color appearance problems. Other difficulties remain, however, such as clothing, the many different contexts in which the same body part can appear, and the high dimensionality of the articulated motion. The work in [25] presented a solution to human pose estimation with a machine learning approach: a classifier was trained with a very large database from simple offset features capturing depth differences between near pixels. Invariance to clothing, body types, and part appearances was therefore learned. This resulted in a very robust solution with previously unseen performance levels, assuming that depth is available.

In psychological studies, however, there is still a clear need for processing and analyzing in RGB-only images: most of the historical material and even newer studies [34] use traditional video, as psychology tends to be a discipline in which technological changes take time to be widely adopted. In

RGB images, the approach of [25] is not directly transferable: simple color differences as features would depend on the background and person appearance. In motion capture, several techniques have been proposed in order to get invariant features from RGB images. HOG-based Body Part Detectors (BPDs) in particular are able to obtain a rough estimate of the pose, later refined with a number of different solutions such as global coherence [30] or symmetry analysis [21]. However, the output of BPDs is often very noisy, with several parts interfering with one another. BPDs are also sensitive to changes in the background.

In this paper, we present a robust method to extract the body pose from sequences of seated conversations, applicable to RGB video data. Our main contributions therefore are: (1) our method provides a high degree of appearance and scale invariance while using only an uncalibrated RGB camera. To this end, we obtain a depth estimate through image motion, given the relationship between both [32]. The scale and spatial context problem is tackled with a single human body detector [9][7], thus avoiding the clutter of using many part detectors. To overcome the lack of information when there is no movement present, we integrate a Kanade-Lucas-Tomasi (KLT) tracker [29]. (2) We evaluate the method's performance with two existing databases and a database of upper body poses displayed in job interviews that we make public, showing that in our scenario it is comparable to that of [25] without using a range camera, and state-of-the-art in the HumanEva [26] and ChaLearn 2011 [1] datasets, used for motion capture system evaluation.

2. RELATED WORK

As explained, given all the challenges that recovering human pose generates, the single camera approach is considered the most elusive. The literature of monocular motion capture can be organized in many ways, but given the latest developments in the field, we classify it into methods that make use of BPDs and those who use other techniques.

Non BPD-based methods: The first BPD-based methods [8] were largely appearance-dependent, leaving the need for more robust solutions with alternative approaches. In [28], biological motion analysis is performed with graphs and motion capture data in order to infer and detect human movements. Using a similar idea, [15] applies optical flow in constrained situations to recover the position of the body limbs. In [12], a strong body model prior and a tracking method are coupled with optical flow recognition for biological motion perception and 3D lifting. The work in [2] also uses strong motion priors to get 3D from 2D images by mapping silhouette descriptors to body pose configurations. Doing so, no explicit body model is required. Instead of a strong motion prior, [3] uses a detailed upper body mesh model coupled with contours and optical flow to recover the pose. Finally, in [24] a range camera is used, and optical flow distinguishes the limbs in moments where depth fails to output reliable geodesic extrema.

BPD-based methods: A number of HOG-based methods have arisen, solving long-standing problems like automatic initialization. In [30] the co-occurrence relations and spatial tree-structured relationships between part detectors are modeled to improve global coherence. Coherence can also be improved with HOG co-dependent body-part Random Forests regressors [6], whereas [21] relies on symmetry analysis. In [11], HOG detectors are improved with skin color, contours, and contextual cues. In [27], 2D and 3D inference are simultaneously done with a generative Bayesian framework and discriminative HOG-based BPDs. In [33], 3D poses are retrieved in unconstrained videos through BPDs, action classification, and pose regression with spatiotemporal features. Finally, in [22] global and local pose cues are included and a convex objective and joint training for mode selection and pose estimation is used to improve the state-of-art performance-to-computing time ratio.

Closest techniques to ours: In [4], upper body motion capture is performed for language sign classification in long videos. Arm detections are used in order to disambiguate difficult hand detections. The torso shape is also inferred from a series of heuristics. However, this method assumes that the hands are generally visible, and the clothing is not problematic. In [23], optical flow and color are used to detect hands, with body part detectors being one of the image features, while [10] relies on optical flow and a precise 2D silhouette body model. In [35] a hand detector is trained with optical flow to then interpolate between correct guesses. However, they use image contours, which are prone to errors in certain situations. In [16] a hand detector is also used, but includes appearance-dependent features such as skin color. To conclude, in [13] multiple body part detectors are trained with Random Forests and per-pixel HOG features, which still inherits some of the appearance-dependence problems that edges entail.

In summary, the current tendencies for monocular RGB motion capture can be grouped into using multiple BPDs (where problems like overlap, interference, or background-appearance dependency arise), the use of strong motion models, strong body models, appearance-relying methods with a torso detector. In contrast, in the present work we combine a single torso detector with largely appearance-invariant features, while improving the best performance-to-computing time ratio of the state-of-the-art.

3. OVERVIEW

An overview of our method can be seen in Figure 2. Given two input consecutive frames $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$, we compute the dense optical flow $\mathbf{I}_{OF}$ [5] and the subject torso bounding box $\vec{b} = [\vec{u_o}, b_w, b_h]$, where $\vec{u_o}$ is the top left pixel of the bounding box, b_w is its width and b_h is its height. We define a 4-channel (4-C) image as $\mathbf{I}_C = [\mathbf{I}_{OF}, \mathbf{I}_{bw}, \mathbf{I}_{bh}]$, where $\mathbf{I}_{bw}$ and $\mathbf{I}_{bh}$ images derived from the torso detection, that aim to give spatial context (see Section 4.1 for details). We then extract per-pixel offset features $\vec{u_\delta}$ from $\mathbf{I}_C$ from a training set (the only prior used) in a similar fashion to [25]. In order to predict the body part classification label image $\mathbf{I}_L$, a Random Forest classifier is used. The label image $\mathbf{I}_L$ and its associated confidence scores $\mathbf{I}_L^{'}$ are used to train a Random Forest regressor that outputs the final body configuration.

4. OBSERVATION SYSTEM

4.1 Largely appearance-invariant features

As explained in Section 3, we aim to extract a series of features from the image that encode as much information as possible while maintaining a high appearance-invariance: they should be robust to clothing and skin color. Thanks to recent advances, dense optical flow and upper body detectors are good candidates. Therefore, we compose a 4-C image

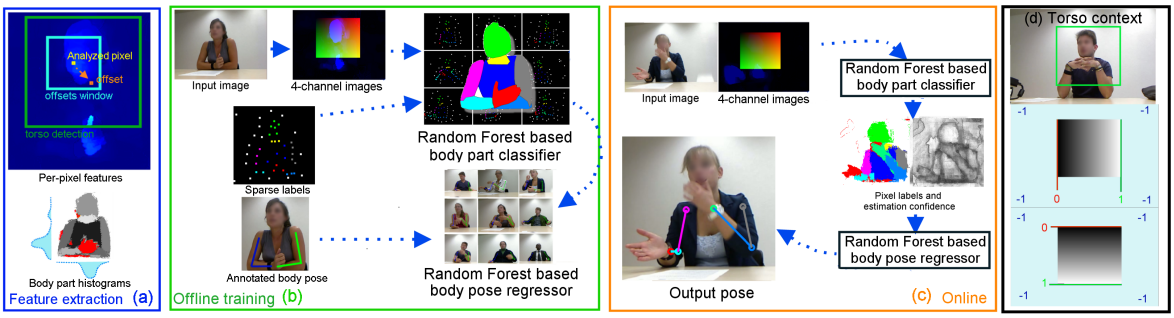


Figure 2: Pipeline of our proposal: (a) Image feature extraction, from the original RGB data and set of body part labels. (b) Offline training of the body part classifier and pose regressor, from sparsely labeled data. (c) Online usage: once the 4-C features have been computed, they are fed into the classifier. In turn, its output is used as input for the regressor, which estimates the body joints configuration. (d) Torso context: horizontal (center) and vertical (bottom) context images $\mathbf{I}_{bw}$ and $\mathbf{I}_{bh}$. Best viewed in PDF.

that merges the information that those low-level features provide. These channels are:

Optical flow modulus $\mathbf{I}_{OF,\rho}$: as mentioned in the introduction, depth and motion in an image are closely related. Therefore, the magnitude of the optical flow vector is a strong cue for positioning the body pose.

Optical flow orientation $\mathbf{I}_{OF,\phi}$: we hypothesize that in certain situations, such as when the hands move close to each other, the optical flow orientation is a useful cue in order to differentiate them.

Torso vertical context $\mathbf{I}_{bh}$: We use the torso detector output to place and estimate the size of the person in the image, adding contextual information. Given the bounding box $\vec{b}$, we encode relative height in the image $\mathbf{I}_{bh}$ (see Figure 2d). In $\mathbf{I}_{bh}$ pixels range from 0 to 1, from the bottom to the top within the bounding box, and are set to -1 outside the bounding box.

Torso horizontal context $\mathbf{I}_{bw}$: Analogous to $\mathbf{I}_{bh}$, while providing horizontal information.

4.2 Body part classification

At this point, highly appearance-invariant images $\mathbf{I}_C$ are obtained. Similar to [25], we use an offset sampling idea and a Random Forest classifier in order to associate every pixel with a body part label. A set $\mathcal{U}$ of pixel offset features is built: $\mathcal{U} = \{\vec{u_{\delta i}}\}_{i=1}^{n_\delta} = \{(u_{\delta i}, v_{\delta i})\}_{i=1}^{n_\delta}$. For a given pixel $\vec{u}$, the feature response is computed with feature parameters $\vec{u_{\delta i}}$ that describe a number of n_δ 2D pixel offsets $(u_{\delta i}, v_{\delta i})$. In [25], features are normalized with the distance to the camera in order to make them depth-invariant. In our case however it is not possible, as we use RGB only images. As a proxy for size of the person, we use the height of the torso bounding box b_h. We then extract the per-pixel features from each channel in a different way.

Optical flow modulus: both the offset distance and optical flow value are normalized with b_h, since motions that take place far from the camera result in a lower optical flow modulus value. Let $L(\vec{u}, \vec{u_{\delta i}}, \mathbf{I}^1)$ be a lookup function that returns the feature associated with pixel $\vec{u}$, given a single-channel image $\mathbf{I}^1$ and an offset $\vec{u_{\delta i}}$. An optical flow modulus feature becomes:

$$f_{OF,\rho}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{OF,\rho})\frac{1}{b_h} - \mathbf{I}_{OF,\rho}(\vec{u})\frac{1}{b_h} \quad (1)$$

It encodes the speed difference between pixel $\vec{u}$ and its associated offset $\vec{u_{\delta i}}$, after normalizing the offset distance and image speed with the torso size b_h.

Optical flow direction: in order to better differentiate parts of the body moving with similar speed modulus, but with different orientations, we also take the optical flow direction into account. For each pixel, we compute the direction similarity in relation to its offset features:

$$f_{OF,\phi}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{OF,\phi}) - \mathbf{I}_{OF,\phi}(\vec{u}) \quad (2)$$

Therefore, given the optical flow angle for pixel $\vec{u}$, the angle difference respective to the offset is computed. The discontinuity between 0 and 360 degrees is addressed so that the angle difference is less than or equal to 180 degrees.

Position relative to the torso: the feature is obtained in an identical way to that of the optical flow direction:

$$f_{bh}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{bh}) - \mathbf{I}_{bh}(\vec{u}) \quad (3)$$

$$f_{bw}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{bw}) - \mathbf{I}_{bw}(\vec{u}) \quad (4)$$

where equations $f_{bh}(\vec{u}|\vec{u_{\delta i}})$ and $f_{bw}(\vec{u}|\vec{u_{\delta i}})$ correspond to the vertical and horizontal context, respectively. With this approximation, each feature gives an idea of where the pixel is positioned in relation to the main portion of the torso.

The feature vector for classification $\mathcal{X}_{tr,class}$ is formed by concatenating all the features, providing information about the speed, direction, and position relative to the torso of every pixel of the image, hence forming a rich representation of human motion. We demonstrate its capabilities in the results section.

4.2.1 *Training the forest for classification*

A subset of data from the real job interview dataset is annotated in order to serve as a training set (see Section 7). For classification, annotations consist of a number of manually-labeled pixels in the image, in which the nature of each label corresponds to a given body part. We then compute the previously introduced per-pixel offset features for each labeled pixel in the training subset. As our method depends on the amount of movement in the image, we discard the pixels with low flow modulus. We then train a Random Forest (of depth 20) with the extracted pixel offset features and the associated body part labels. Given an unseen 4-

C image, the classifier outputs the per-pixel predicted body part and an associated confidence score (see Figure 2).

4.3 Body pose regression

The next step is to obtain the final body configuration $\mathcal{P}_o$ (defined as the pixel position in the image for every joint) from the output of the body part classifier: $\mathcal{P}_o = \Omega(\mathbf{I}_L, \mathbf{I}_L^{'}, \beta)$, where Ω is the regression model with β parameters. Therefore, we use a Random Forest regressor that takes information extracted from an image of densely classified body parts $\mathbf{I}_L$ and classification scores $\mathbf{I}_L^{'}$, resulting in the regression training set $\mathcal{X}_{tr,reg}$, through a process described below.

4.3.1 *Obtain images $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$ to train the regressor*

In order to train the classifier, sparse labels are used. This responds to two reasons. First, to reduce annotation time: it can be reduced to less than half of the time required for dense labeling. Second, to force the classifier to generalize: as during training only a few (∼100) pixels per image are labeled, the trained forest is shown later the same training images, obtaining the classification output for every pixel in the image. In this process, the provided classification scores $\mathbf{I}_L^{'}$ picture more realistically the confidence that the classifier will have in unseen images. This property is important when training the body pose regressor, as it is forced to learn from the mistakes that the body part classifier makes.

4.3.2 *Body part histograms from $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$*

In order to capture more explicitly the characteristics of the predicted body parts placement in the image, we propose the use of vertical and horizontal per-class histograms of $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$. We build three sets of histograms. The first one, $\mathcal{M}_l$, measures the frequency in which each body part appears in the vertical and horizontal axis of the image, and it is defined as $\mathcal{M}_l = [\vec{m_{l,v}}, \vec{m_{l,h}}]$, where $\vec{m_{l,v}}$ and $\vec{m_{l,h}}$ are the vertical and horizontal histograms. The second one, $\mathcal{M}_l^{'}$ adds the scores $\mathbf{I}_L^{'}$ in the pixels of the image that belong to the relevant body part, along the vertical and horizontal axis of the image, and it is defined as $\mathcal{M}_l^{'} = [\vec{m_{l,v}^{'}}, \vec{m_{l,h}^{'}}]$. The third one, $\mathcal{M}_l^{''}$, is the product of the previous two:

$$\mathcal{M}_l^{''} = [\vec{m_{l,v}} \vec{m_{l,v}^{'}}, \vec{m_{l,h}} \vec{m_{l,h}^{'}}] \quad (5)$$

This results in the body part frequency histogram $\mathcal{M}_l$ to be weighted with its associated confidence scores $\mathcal{M}_l^{'}$. The effect can be seen in Figure 3: the resulting histogram $\mathcal{M}_l^{''}$ shows better where the most confident predictions are located in the image, for a given body part. In order to reduce the dimensionality of the histograms, the images $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$ are down-sampled to a resolution of 128x96 pixels. Therefore, a given vertical histogram becomes 96-dimensional, and a horizontal histogram becomes 128-dimensional. Since there are 10 different body parts classes, 10 different set of histograms $\mathcal{M}_l^{''}$ are obtained (one for each body part). This results in a 2240-dimensional feature vector $\mathcal{X}_{tr,reg}$ for regression $\mathcal{X}_{tr,reg} = \{\mathcal{M}_{l,i}^{''}\}_{i=1}^{n_j}$, where n_j is the number of body parts (classes). For each feature vector $\mathcal{X}_{tr,reg}$ there is an associated annotated body pose, consisting in the pixel position of 6 joints: shoulders, elbows and wrists. It is therefore 12-dimensional. Both the feature vector and the labels are used to train the regression model Ω.

As a summary, in order to obtain the body pose from a unseen pair of $\mathbf{I}_t$, $\mathbf{I}_{t-1}$ images, the associated 4-C image $\mathbf{I}_C$ is first composed, and then input into the body part classifier. The resulting densely-labeled image and confidence scores ($\mathbf{I}_L^{'}$ and $\mathbf{I}_L$) are fed into the body pose regressor, which outputs the predicted final pose configuration $\mathcal{P}_o$ as pixel positions of every joint.

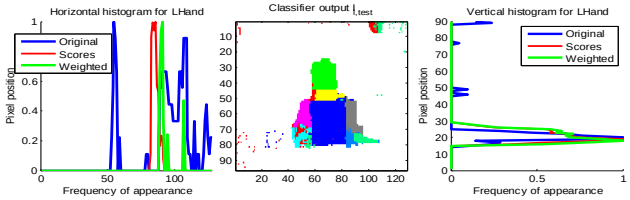


Figure 3: Body part histogram example. Left and right: $\vec{m_{l,v}}$ and $\vec{m_{l,h}}$ for the left hand. Center: classifier output. The confidence weighting helps to maintain the detection peak in the right position. Best viewed in color.

5. TRACKING

In order to reliably obtain the body pose in every situation, we propose a tracking method where hands are tracked using the KLT framework when the body part classifier is not reliable. We then impose temporal smoothness to avoid sudden changes in the pose.

Tracking the hands. The main drawback of our proposal is the need for movement in the image, as the classifier needs optical flow measurements in order to classify different body parts. We extend the pose retrieval framework by adding a KLT tracker, based on image features. KLT trackers work better with slow motion, as motion blur and quick appearance changes become a problem when obtaining good features to track. On the other hand, it has shown to be very reliable when small motions are present. Since our pose detection method is most confident when movement is present, the KLT functions in a natural, complementary way with the pose regressor.

Taking this into account, a detection followed by tracking framework is proposed to obtain the body pose along a whole video. As finding the hands' position removes a high degree of uncertainty in the pose [17], we employ this approach only in them. When the scores image $\mathbf{I}_L^{'}$ falls under a manually set threshold for a hand region, the detection is deemed unreliable, and the last reliable detection is used to initialize a KLT tracker, that takes control of the hand position until a new reliable detection is found, usually when the hand starts to move again. An example can be seen in Figure 4. Since both hands are considered separately, two KLT trackers are used, one for each hand.

Stabilizing the body pose. At this point, a body pose configuration $\mathcal{P}_o$ is available. Given that the ultimate goal of our method is to serve as information to later analyze nonverbal communication, a further post-processing step is needed. When a subject stands still, ideally the output pose would remain perfectly still too. However, as the system involves a high degree of tracking by detection, $\mathcal{P}_o$ slightly fluctuates along time.

In order to compensate this effect, a simple yet effective approach is followed. If the joints stay within some radius for a certain period then they are frozen in that position. Especifically, for each body joint a small pixel radius κ_m is defined. If the predicted body joint falls under the radius

longer than a small pre-defined time t_m, then the joint is defined as the center of the static circle that is configured with κ_m. The timing parameter t_m is necessary in order to avoid the discretization of continuous movements, as they would evolve in κ_m steps otherwise.

6. DATASETS

6.1 Upper body pose interview database

We construct a case-relevant database (that we make public) from a set of one-shot, non-consecutive color and depth images extracted from a series of real job interview videos (kindly shared by [19, 20]), containing 34 different subjects (8 male, 26 female), as Figure 7 shows. The interviewee is sitting in front of a table, and a Kinect device is pointing frontally at him/her, therefore only the upper-body is visible. As the images have been extracted from real job interviews, the appearance, clothing, or movement of the participants are not restricted or acted in any way. We selected 1420 frames with a resolution of 640x480 both for RGB and depth, in which the subject is moving at least a part of the body. As shown by previous works with Random Forests, a large amount of training data is required, therefore we distributed the frames as 1100 for training and 320 for testing. Even if the amount of data is significatively lower that one of [25], given the similarity of the underlying methods, it still allows comparison in generalization capabilities. In addition, as only the upper-body is considered, the degrees of freedom of the body configuration is much lower. The data will be made publicly available with blurred faces due to privacy reasons.

6.1.1 Annotation

Test images ground-truth is obtained by manually labeling every pixel of the testing images with 10 labels: right and left hands, right and left forearms, right and left arms, head, torso, neck and background (see Figure 2). For the training images, manual sparse labeling has been followed. That is, we only label a few (around 100) pixels per image, in the parts that carry less uncertainty. For example, if there is a self-occlusion, only the pixels in which the different parts of the body are clearly distinguished are labeled. Manual sparse labeling carries two main advantages: first, the labeling time is greatly reduced; secondly, it allows us to choose the parts of the image that better represent each part. With the aid of a purpose-built script, labeling takes an average of only 25 seconds per image, with a speed close to 5 pixels per second. In contrast, dense labeling would take almost 1 minute per frame.

As seen in Figure 2, if enough data has been sparsely labeled, it is possible to get an approximate dense labeling by using the method in [25] on the sparse labels, which can be manually corrected later. Also, we effectively double the amount of training data by mirroring each image, both during classification and regression. Finally, the groundtruth for the regression task is obtained by annotating the joint positions of the wrist, elbow and shoulder position in every image.

6.2 Additional datasets

In order to validate our methods, show generalization of performance, and allow for comparisons, we test our proposal in two publicly-available datasets. We choose the subsets of experiments that are most relevant to our scenario: similar upper body movements to those that are found in conversations.

ChaLearn 2011 contains 437 non-consecutive, 320×240 color and depth frames, in which body joints have been manually annotated. The environment is uncontrolled, and different backgrounds, high variance of poses, clothing, positioning and lighting appear. In some of the frames, there is no movement at all.

HumanEva-I contains 7 calibrated video sequences (4 grayscale and 3 color) that are synchronized with 3D body poses obtained from a motion capture system. The database videos contain 4 subjects (S1-S4) performing a 6 common actions (e.g. walking, jogging, gesturing, etc.). The dataset contains training, validation and testing sets. We chose sequences 'gestures' and 'box' sequences of S1 and S3 for evaluation.

7. RESULTS

We evaluate the effects of different parameters in our system's performance, and compare it with the current state-of-the-art. For the one-shot-detection part of the approach, we define two experiments: one for classification and one for regression. For classification, we compare the output of our method to annotations of every body part. The result is given in per-pixel accuracy for each class. For regression, we compare the output pose that we get with manual annotations of the database we used, measuring joint detection rate. A joint is defined as detected if the predicted point falls within a given distance threshold.

7.1 Classification

The performance is evaluated in our job interview database. We train both our algorithm and [25] with the same number of images. As our method depends on the amount of movement in the image, we discard the pixels that fall below a motion threshold. This results in fewer but reliable data points. In total, more than 110k pixels were used for depth training, and 88k for optical flow training. The same classifier parameters were used in both cases: 75 trees, an offset window of 250 pixels and 700 features. We generate 320x240 classified images in order to have a reasonable resolution while maintaining a competitive processing time. The results can be seen in Figure 4 for the different cases.

Using depth, a **67.7%** accuracy is achieved for the pixels associated with the person, which is consistent with the previous findings of [25], when taking into account the much lower training information. Over the whole image, the accuracy is **92.3%**. Using optical flow and torso detections, the accuracy for the person is **63%**, and **87.7%** for the whole image. This shows that our method can achieve similar results to depth when there is body movement. An interesting finding is that our method outperforms [25] in hand detection. As the reviewed literature shows, hand position is a very good proxy in order to infer the rest of the body pose. The usefulness of optical flow in hand and arm detection is confirmed, as when combining depth and optical flow modulus, the body accuracy increases to **70%**, with the whole image at **92.2%**.

Sensitivity to parameters results can be seen in Figure 5. When considering offset window size before scale normalization, it is found that a size of 200x200 improves the body-related detections, but introduces more background noise than 250x250 sizes, making the regression system more prone

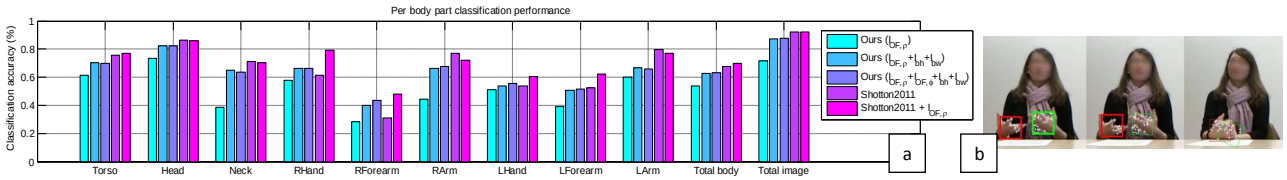


Figure 4: Left: **Results of our method versus [25]**, when trained with the same number of images: Reported as classification rates for every body segment. Performance is asymmetrical as the camera is slightly off-center. Right: **Integration with the KLT tracker example**. Reliable detections are represented with a thick box. The good features to track are shown as white crosses.

to errors. Surprisingly, using a very low number of offset features did not cause big performance drops. Already with 150 pixel offset features, very competitive results are obtained, as a result of the rich contextual information that the torso detector provides. As for the number of trees used, it is found that after 5 trees, there is only a slight performance increase to be found. Finally, when considering the amount of training data, competitive results can already be obtained with 300 training images (effectively 600, since they are mirrored). This highlights the generalization capabilities of the proposed features.

7.2 Regression

Performance is measured with two different datasets: our upper-body interview dataset, and ChaLearn 2011, which is non-conversational, but allows to show generalization of performance. Rates of correct detected parts can be found in Figures 5. We found 40 pixels to be the limit of a reliable guess while using 640x480 images.

Job interview dataset. Our method performs similarly to [25], and outperforms [9] and [22]. The latter work had been trained with their FLIC database, and is considered as the best method in terms of performance to processing time ratio. Wrists are the hardest body part to detect, and our method achieves close to **20%** higher performance than [22]. When compared against [25], our method is less than **10%** behind. Using the classification weights scores, accuracy is improved by almost **5%** for the hands. In Figure 7 some qualitative results can be found.

ChaLearn 2011 We perform leave-one-out after dividing the data in 10 arbitrary groups. As it can be seen in Figure 5, the trends shown in our job interview dataset are reproduced when there is enough movement available. The gap to [25] appears larger due to several factors: (a) less training data: only 57% of the total labeled points are used in our approach as we only use points that contain movement information in order to train the forest. This also gives an idea of the little movement information present in the dataset. As Figure 7 top right shows, in the left hand there is a higher amount of movement, greatly reducing the gap with [25]. (b) Some subjects move out of frame (the head or one arm is not visible), making it a hurdle for the torso detection to be correctly placed, and reducing the accuracy of shoulders and elbows. (c) In some cases the subject casts shadows in the background wall, producing spurious optical flow detections. Despite this, we found that in some instances the context provided by the torso detector is able to filter errors out. In any case, our method performs clearly better than the RGB-only baselines, and given the challenging conditions, remarkably close to [25] when there is movement present. Also, it shows that our system performs well with different body scales.

7.3 Tracking

In order to evaluate the tracking proposal, we define three experiments. The **experiment #1** evaluates hand positioning, as it is a very good proxy for the global pose [17]. During two minutes of video from 4 subjects, the hand position was manually annotated. The **experiment #2** follows the same approach, but containing a very challenging 30 second sequence (see column 3 of Figure 7 left), as she has no sleeves (therefore a lot of skin exposed, which produce appearance-dependence situations when using a skin segmentation scheme), moves her hair which is a similar color to that of the skin, and displays a series of unusual movements (such as the shoulders being closer to the camera than the hands at some points). Finally, the **experiment #3** uses the HumanEva I database.

The baseline for the experiments #1 and #2 is a state-of-the-art hand detection measure [31]. Hand saliency is defined as $\mathbf{I}_{H,t} = \mathbf{I}_{S,t} \cdot \mathbf{I}_{F,t} \cdot (\kappa_1 \mathbf{I}_{OF,\rho,t} + \kappa_2 \mathbf{I}_{D,t})$, where $\mathbf{I}_{S,t}$ is the skin segmentation, $\mathbf{I}_{F,t}$ is the face region subtraction, $\mathbf{I}_{D,t}$ is the distance to the camera, and κ_1 and κ_2 are manually set constants. For comparison purposes between RGB-only methods, we use the same saliency measure without the addition of $\mathbf{I}_{D,t}$, denoted as $\mathbf{I}'_{H,t}$. We apply a best path tracking scheme to the local maxima of $\mathbf{I}_{H,t}$ and $\mathbf{I}'_{H,t}$.

The results of our tracking method applied for hand position can be seen in Figure 6. In experiment #1, the hand detection rate is **78.6%**, marginally higher than the baseline obtained with $\mathbf{I}_{H,t}$ (**78.2%**), even though no depth information is used in our approach. When compared to the performance of $\mathbf{I}'_{H,t}$, our framework shows a clear increase of performance. In the challenging experiment #2, we found that our approach significantly increases the performance gap with respect to the $\mathbf{I}_{H,t}$ baseline, which uses depth, obtaining a detection rate of **66%**, compared to **31%**. This clearly shows that our method does not simply label body parts that move the fastest as hands, but rather takes the actual shape of the body part movements into account in the learning process. It also carries the extra advantage of being able to explicitly distinguish between right and left hands.

As Figure 6 shows, we obtain state-of-the-art performance in HumanEva I (in [18] an average of 12.6 pixel error is reported). The largest errors occur when the performed pose substantially differs to those contained in the training set. Given the very few data samples that HumanEva I makes available for training (we use an average of 46 sparsely labeled frames per sequence), and the fact that our method requires large training sets [25], we find the performance very encouraging overall. See Figure 7 for qualitative results.

7.4 Computing time

Computing time is key in psychological studies, given the large amount of data. In [22], there appears an analysis

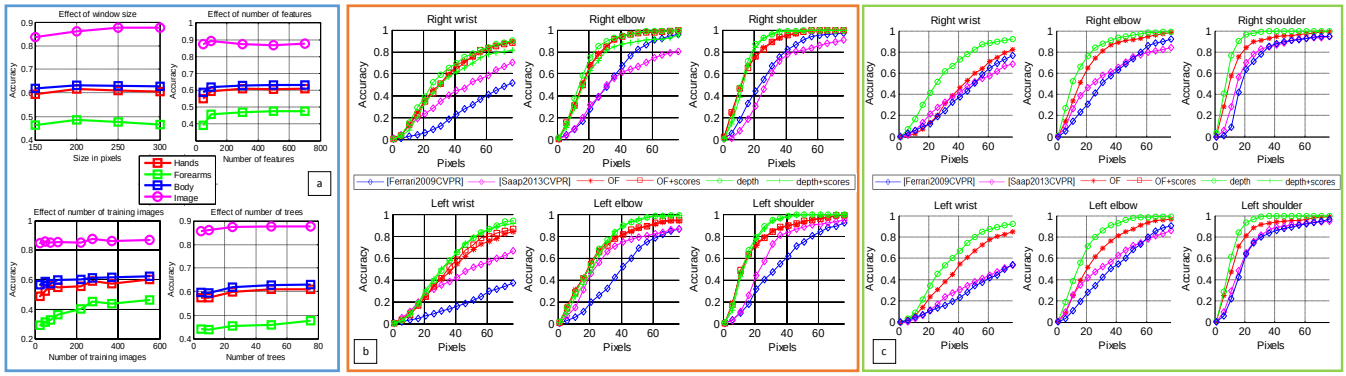
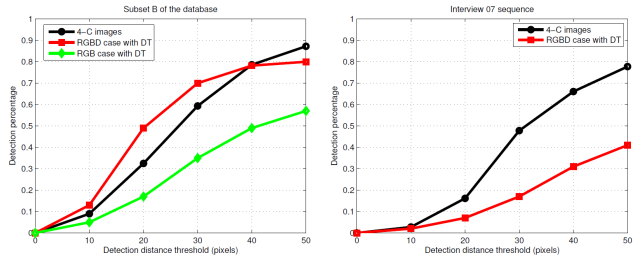


Figure 5: a: Parameter sensitivity of the classifier. b: Accuracy of the regressor in our job database, compared against [9] (blue) and [22] (magenta). c: Results in ChaLearn2011, compared against [9] (blue) and [22] (magenta). Best viewed in PDF in high zoom.



Subject	Action	Error (pixels)
S1	Gestures	5.0
S1	Box	13.2
S3	Gestures	11.3
S3	Box	21.3

Figure 6: Tracking results. Left graph: first experiment. Right graph, using a specially challenging sequence. Table: Average error in HumanEva. Best viewed in color.

of the state-of-the-art performance versus computing time, with their work being the best placed. Using the code they provide, our database is computed on a laptop with an Intel i7 processor in an average of 5.18 secs per frame (standard deviation 0.16 secs).

In our case, assuming that pre-computed optical flow is available (it can be comfortably processed in real time with modern GPUs), the average processing time from input features to body pose is **1.59 secs** per frame (standard deviation 0.11 secs), using the same hardware. Our per-pixel feature retrieval implemented in Matlab takes most of the running time. Since regression needs the output of the classifier with a resolution of 128x96 pixels in order to build the histograms, we obtain features for every fifth pixel of the 4-C composed image. If only classification results are needed, we obtain comparable processing times to [22] by using a resolution of 320x240 for I_l, recording a mean of 5.05 secs and (standard deviation 0.14 secs). This offset feature-based approach is implemented on an XBox 360 in [25] at 200 fps. Finally, an average of an extra **0.027 secs** per frame is required in order to track the hands with the KLT approach.

Limitations: Our method requires a static camera and static background, but optical flow based methods in the literature have shown to overcome that problem by tracking background features. As the torso bounding box misplacements are one of the main sources of error, our approach can highly benefit from an elaborated torso bounding box tracking technique.

8. CONCLUSIONS

We proposed a fast, largely appearance-invariant method for upper body monocular motion capture of people engaged in conversation, which integrates detection and tracking. Detection was achieved through optical flow and body detectors, providing a proxy for depth information, visual context, and scale. This information was used to classify body parts in the image with a Random Forests classifier. The classification output and per-pixel confidence was later used to build per body part image histograms, and were fed to a regressor in order to infer the body pose. The integration of a KLT tracker allowed to follow the body pose when there are no reliable detections, thus resulting in a complementary framework.

We evaluated our method with three different datasets, showing very close performance to that of the best depth-based method, while using only monocular information. We also clearly outperform the state-of-the art in the ratio accuracy to processing time. Our method is therefore attractive to process video data in typical psychology lab studies, where depth data is not yet available. Our database of static upper-body poses in interviews will be made public, providing a reliable benchmark for real-world performance.

9. ACKNOWLEDGMENTS

This research was funded by SNSF SONVB and UBImpressed projects, the Spanish Ministry of Economy and Competitiveness under project SPACES-UAH (TIN2013-47630-C2-1-R) and the University of Alcala FPI program.

10. REFERENCES

[1] Chalearn gesture dataset (cgd2011), chalearn, california, 2011. copyright (c) chalearn - 2011. http://gesture.chalearn.org/data.

[2] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE PAMI*, 28(1), 2006.

[3] T. Brox, B. Rosenhahn, and D. Cremers. Contours, optic flow, and prior knowledge: cues for capturing 3D human motion in videos. In *Human Motion - Understanding, Modeling, Capture, and Animation*. 2007.

[4] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *IJCV*, 95(2), 2011.

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

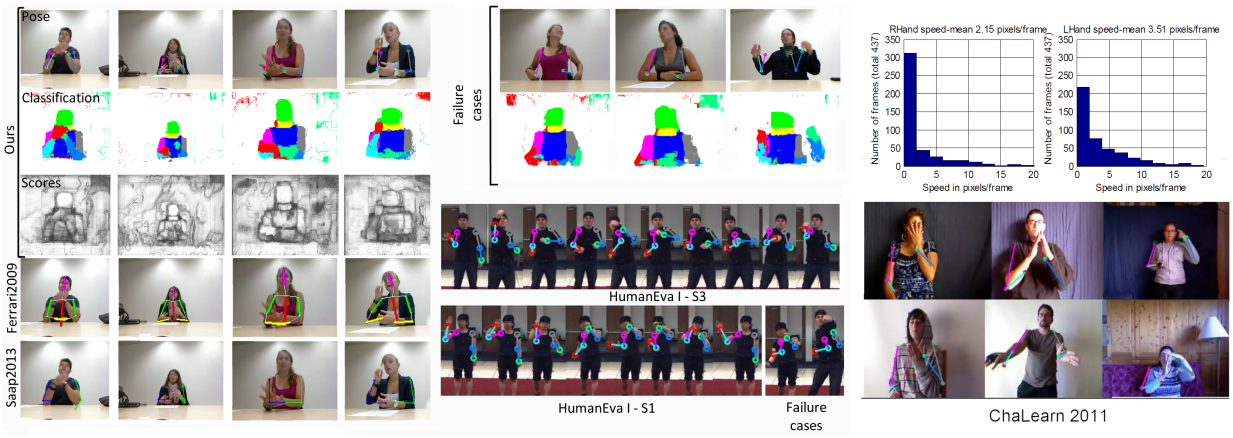


Figure 7: Qualitative results. Left: comparison with the baselines [9] and [22]. Top center: failure cases. From left to right: untrained pose, not enough movement information, torso detection failure. Bottom center: HumanEva I results for S1 and S3. Failure cases, from left to right: untrained pose, torso detection failure. Top right: speed histograms for each hand. Bottom right: examples of obtained poses in ChaLearn 2011. Best viewed in PDF in high zoom.

[6] M. Dantone, J. Gall, C. Leistner, , and L. V. Gool. Human pose estimation using body parts dependent joint regressors. In *IEEE CVPR*, 2013.

[7] P. Dollar, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *IEEE ECCV*, 2012.

[8] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *IEEE CVPR*, 2000.

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: Retrieving people using their pose. In *IEEE CVPR*, 2009.

[10] K. Fragkiadaki, H. Hu, and J. Shi. Pose from flow and flow from pose. In *IEEE CVPR*, 2013.

[11] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *IEEE CVPR*, 2013.

[12] N. R. Howe. Flow lookup and biological motion perception. In *International Conference on Image Processing (ICIP)*, 2005.

[13] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *British Machine Vision Conference (BMVC)*, 2013.

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

[15] D. Kulic, D. Lee, and Y. Nakamura. Whole body motion primitive segmentation from monocular video. In *IEEE ICRA*, 2009.

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

[17] E. Marinoiu, D. Papava, and C. Sminchisescu. Pictorial Human Spaces. How Well do Humans Perceive a 3D Articulated Pose? In *IEEE ICCV*, 2013.

[18] V. Morariu, D. Harwood, and L. Davis. Tracking people's hands and feet using mixed network and/or search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35, 2013.

[19] L. Nguyen, A. Marcos, M. Marron, and D. Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *ACM ICMI*, 2013.

[20] L. S. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, 2014.

[21] V. Ramakrishna, T. Kanade, and Y. Sheikh. Tracking human pose by tracking symmetric parts. In *IEEE CVPR*, 2013.

[22] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *IEEE CVPR*, 2013.

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

[24] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab. Estimating human 3d pose from time-of-flight images based on geodesic distances and optical flow. In *IEEE FG*, 2011.

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

[26] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1), 2010.

[27] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. In *IEEE CVPR*, 2013.

[28] Y. Song, L. Goncalves, and P. Perona. Learning probabilistic structure for human motion detection. In *IEEE CVPR*, 2001.

[29] C. Tomasi and T. Kanade. Detection and tracking of point features. *IJCV*, 1991.

[30] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE CVPR*, 2011.

[31] Y. Yin and R. Davis. Gesture spotting and recognition using salience detection and concatenated hidden markov models. In *ACM ICMI*, 2013.

[32] A. Yoonessi and C. L. Baker. Contribution of motion parallax to segmentation and depth perception. *Journal of Vision*, 11, 2011.

[33] T.-H. Yu, T.-K. Kim, and R. Cipolla. Unconstrained monocular 3d human pose estimation by action detection and cross-modality regression forest. In *IEEE CVPR*, 2013.

[34] X. Yu, S. Zhang, Y. Yu, N. Dunbar, M. Jensen, J. Burgoon, and D. Metaxas. Automated analysis of interactional synchrony using robust facial tracking and expression recognition. In *FG Workshops*, 2013.

[35] S. Zuffi, J. Romero, C. Schmid, and M. J. Black. Estimating human pose with flowing puppets. In *IEEE ICCV*, 2013.