Three-dimensional Reconstruction of Human Interactions

Mihai Fieraru\textsuperscript{1} Mihai Zanfir\textsuperscript{1} Elisabeta Oneata\textsuperscript{1}
Alin-Ionut Popa\textsuperscript{1} Vlad Olaru\textsuperscript{1} Cristian Sminchisescu\textsuperscript{2,1}

\textsuperscript{1}Institute of Mathematics of the Romanian Academy, \textsuperscript{2}Lund University
\textsuperscript{1}\{firstname.lastname\}@imar.ro, \textsuperscript{2}cristian.sminchisescu@math.lth.se

Abstract

Understanding 3d human interactions is fundamental for fine grained scene analysis and behavioural modeling. However, most of the existing models focus on analyzing a single person in isolation, and those who process several people focus largely on resolving multi-person data association, rather than inferring interactions. This may lead to incorrect, lifeless 3d estimates, that miss the subtle human contact aspects--the essence of the event--and are of little use for detailed behavioral understanding. This paper addresses such issues and makes several contributions: (1) we introduce models for interaction signature estimation (ISP) encompassing contact detection, segmentation, and 3d contact signature prediction; (2) we show how such components can be leveraged in order to produce augmented losses that ensure contact consistency during 3d reconstruction; (3) we construct several large datasets for learning and evaluating 3d contact prediction and reconstruction methods; specifically, we introduce CHI3D, a lab-based accurate 3d motion capture dataset with 631 sequences containing 2,525 contact events, 728, 664 ground truth 3d poses, as well as FlickrCI3D, a dataset of 11,216 images, with 14,081 processed pairs of people, and 81,233 facet-level surface correspondences within 138,213 selected contact regions. Finally, (4) we present models and baselines to illustrate how contact estimation supports meaningful 3d reconstruction where essential interactions are captured. Models and data are made available for research purposes at \url{http://vision.imar.ro/ci3d}.

1. Introduction

Human sensing has recently seen a revival\cite{34, 28, 24, 48} due to advances in large-scale deep learning architectures, powerful 3d kinematic and statistical shape models\cite{23, 14, 32}, as well as large scale 2d and 3d annotated datasets\cite{20, 1, 13, 25}. While considerable progress has been achieved in localizing multiple humans in images, or reconstructing 3d humans in isolation, in a person-centred frame, little work has focused on inferring the pose and placement of multiple people in a three-dimensional, scene-centered coordinate system.

Moreover, the few approaches that have pursued such goals recently\cite{29, 50, 51, 19, 40, 3, 16}, concentrated mostly on the arguably difficult problem of multi-person data association, rather than the more subtle aspects such as close interactions during human contact. This leads to predictions that even when impressive in terms of plausible pose and shape from a distance, miss the essence of the event at close scrutiny, when e.g. two reconstructions fail to capture the contact during a handshake, a tap on the shoulder, or a hug. Such interactions are particularly difficult to resolve as effects compound: on one hand depth and body shape uncertainty could result in compensation by pushing limbs in front or further away from their ground truth position, when inferring 3d from monocular images\cite{39}; on the other hand, partial occlusion and the relatively scarce detail (resolution) for contact areas in images, typical of many human interactions, can make visual evidence inconclusive.

In this paper, we propose a first set of methodological elements to address the reconstruction of interacting humans, in a more principled manner, by relying on recognition, segmentation, mapping, and 3d reconstruction. More precisely, we break down the problem of producing veridical 3d reconstructions of interacting humans into...
contact detection, (b) binary segmentation of contact regions on the corresponding surfaces associated to the interacting people; (c) contact signature prediction to produce estimates of the potential many-to-many correspondence map between regions in contact; and (d) 3d reconstruction under augmented losses built using additional surface contact constraints given a contact signature. To train models and evaluate the techniques we introduce two large datasets: CHI3D, a lab-based 3d motion capture repository containing 631 sequences containing 2,525 contact events, 728,664 ground truth skeletons, as well as FlickrCI3D, a dataset of 11,216 images, with 14,081 processed pairs of people, and 81,233 facet-level surface contact correspondences. In extensive experiments, we evaluate all system components and provide quantitative and qualitative comparisons showing how the proposed approach can capture 3d human interactions realistically.

**Human and Object Interactions.** The 3d reconstruction of multiple people in the context of close interactions was partially addressed in [21, 22], where 3d human skeletons were rigged to mesh surfaces of participants. Scenes were captured using a multi camera setup on a green background. 3d pose estimation and shape modelling were performed using energy-based optimization, taking into account the multi-camera setup, green background separation and temporal consistency. Human interactions or contact were not modeled explicitly beyond non-penetration of mesh surfaces. Yun et al.[49] proposed methods for action classification in scenes with two interacting people using RGB-D data and multiple instance learning. However, their data does not imply physical interaction between subjects and no form of contact is labeled. Hand to hand interaction is studied in [44, 45, 42, 30], where models are optimized using energy minimization with non-penetration constraints but without a contact model. Other methods focus on the interaction between the 3d human shape and the surrounding environment[10, 31, 2, 44, 33, 11, 4], in most cases without a detailed object contact model.

**Psycho-social Studies.** [41] construct a body region map to describe the most likely contact areas for different types of social bonds (e.g. child and parent, siblings, life partners, casual friends, strangers) and conclude that social contact between two individuals varies with emotional bondage. Human close interaction analysis could be important in social studies involving robot assisted therapy for autistic children. [27, 36] record robot assisted therapy sessions and perform extensive analysis over the interactions between the therapist, the children and the robot. Research in this area can impact social group interaction analysis c.f. [37, 6]. A similar study [18] was performed over the dyadic relationship between mother and children.

**Datasets.** Most datasets dedicated to human understanding are centered around single person scenarios[1, 13, 38, 25] and even those that include multiple people [20, 46, 7] do not explicitly model the close interaction between different people. Human interactions have been captured before in classification/recognition datasets, one such example being [9]. The dataset provides action labels for short video sequences (i.e. under 1-2 seconds) collected from YouTube. However, the dataset does not provide detailed contact annotations either in the image or at the level of 3d surfaces, as pursued here.

### 2. Datasets and Annotation Protocols

**FlickrCI3D.** We collect images from the YFCC100M dataset[43], a database containing photos uploaded to Flickr by amateur photographers who share their work under a Creative Commons license. Using [15] to search the dataset, we download images expected to contain scenes with close interactions between people. We either query the dataset using tags generic to the human category, such as "persons", "friends", "men", "women", or using tags related to actions performed by humans in physical contact, such as "dance", "hug", "arrest", "handshake". We run a common multi-person 2d keypoint estimator[5] - to detect the humans in each picture and select all pairs of people whose bounding boxes overlap. We automatically filter out the images with small resolution, where pairs of people are severely occluded or have large scale differences. We refer to this data collection together with the underlying 3d surface contact annotations (described in the sequel) as FlickrCI3D.

**CHI3D.** We also collect a lab-based 3d motion capture dataset, CHI3D (Close Human Interactions 3D), for quantitative evaluation of 3d pose and shape reconstruction. We employ a MoCap system of 10 motion cameras synchronized with 4 additional RGB cameras. We capture short video sequences of 6 human subjects, grouped in 3 pairs, performing close interaction scenarios: grab, handshake, hit, holding hands, hug, kick, posing (for picture) and push. To preserve realism as much as possible, each of the human subjects takes turns on wearing the body markers. To obtain a pseudo-ground truth 3d pose configuration of the person not wearing the markers, we run a 2d keypoint estimator[5] in each of the 4 views and perform robust triangulation.

In total, we collect 631 sequences consisting of 485, 776 pairs of RGB frames and MoCap skeleton configurations. Using triangulation, we manage to reconstruct an additional 242, 888 3d skeleton configurations. The number of pseudo-ground truth 3d skeletons is smaller since the 2d keypoint estimator sometimes fails for people in close proximity, which causes the triangulation to fail as well.
2.1. Annotation Protocol

We next describe the manual annotation pipeline, which consists of two stages. First, the annotators label whether or not two people are in contact. Second, they localize the physical contact for pairs of people annotated as being in contact, by establishing correspondences between two 3D human body meshes.

In both steps, the annotators are presented with a picture and two superimposed 2D skeletons identifying the people of interest. This approach helps clear the identity confusion in crowded scenes or in interactions with high overlap between people.

Contact Classification. Given a scene where the detector identified 2D body poses in close proximity, we identify four scenarios that have to be manually classified: (1) erroneous 2D pose estimations, i.e. the assignment between the estimated skeletons and the people in the image cannot be determined, or at least two estimated limbs have no overlap with the real limbs in the image, (2) certainly no contact between the two people, (3) contact between the two people and (4) uncertain contact between the two people, i.e. both "contact" and "no contact" cases may be possible, but it is ambiguous in the given image. On FlickrCI3D, annotators are instructed to label each pair of people with one of these four classes, which is achieved at an annotation rate of around 500 pairs / hour. By discarding the few pairs of 2D skeletons that are erroneous (8%), the result is a database of 65,457 images, containing 90,167 pairs of people in close proximity, in the following proportions: 18% "contact", 21% "uncertain contact", 61% "no contact". Example images from each of these classes are shown in fig. 2. On CHI3D, annotators are instructed to select only one frame per video sequence where people are in physical contact. Since more information is available, we show annotators all 4 views, as well as the temporal context of the corresponding sequence. This results in a total of 2,524 frames of people in contact.

3D Contact Signature Annotation. When two people are in physical contact, we want to understand where and how they interact by encoding the information on the surfaces of two 3D human meshes.

To this end, we define the facet-level contact signature \( C_{\text{facet}}(I, P_1, P_2) \in \{0, 1\}^{N_{\text{facets}} \times N_{\text{facets}}} \) between two people \( P_1, P_2 \) in image \( I \) as \( C_{\text{facet}}(I, P_1, P_2) = 1 \) if facet \( f_1 \) of the mesh of person \( P_1 \) is in contact with facet \( f_2 \) of the mesh of person \( P_2 \) and \( C_{\text{facet}}(I, P_1, P_2) = 0 \) if they are not in contact.

We also define the facet-level contact segmentation \( S_{\text{facet}}(I, P_1, P_2) \in \{0, 1\}^{N_{\text{facets}} \times 2} \) of the contact of two people \( P_1, P_2 \) in image \( I \) as \( S_{\text{facet}}(I, P_1, P_2) = 1 \) if facet \( f \) of the mesh of person \( P_1 \) is in contact with any other facet of the mesh of the other person, and \( S_{\text{facet}}(I, P_1, P_2) = 0 \) otherwise. Note that the contact segmentation \( S \) can be recovered from the contact signature \( C \).

State of the art body meshes \([47, 32]\) have a large number of surface facets, \( N_{\text{facets}} \approx 20,000 \). Annotating a contact signature with high fidelity in such a huge dimensional space, i.e. \( N_{\text{facets}} \times N_{\text{facets}} \), is both tedious and time-consuming.

An alternative to simplify annotation burden is to first perform segmentation and then establish correspondences only between the contact segments on both surfaces. However, even fine segmentation annotation of the contact on 3D surfaces is cumbersome and requires a high degree of precision and creativity. Instead, we relax the annotation granularity and group the \( N_{\text{facets}} \) facets into a number \( N_{\text{reg}} = 75 \) predefined regions. We guide our grouping strategy by following the anatomical parts of the human body and their symmetries, as seen in fig. 3.

Now, the definitions of region-level contact signature \( C_{\text{reg}}(I, P_1, P_2) \in \{0, 1\}^{N_{\text{reg}} \times N_{\text{reg}}} \) and region-level contact segmentation \( S_{\text{reg}}(I, P_1, P_2) \in \{0, 1\}^{N_{\text{reg}} \times 2} \) follow naturally by considering two regions \( r_1 \) and \( r_2 \) to be in contact if at least one facet from region \( r_1 \) is in contact with at least one facet from region \( r_2 \). With such a setup, segmentation can be performed quickly with a few clicks on the regions in contact.

To support the annotation effort, we implemented a custom 3D annotation interface which displays an image, the superimposed 2D poses of the people of interest, alongside...
Table 1: Annotator consistency as a function of the granularity of surface regions. The task has an underlying ground truth, but it is sometimes hard for annotators to identify it. At 17 and 9 regions partitioning, respectively, there is reasonable consistency. Notice that for CHI3D the consistency is higher as the annotators rely on 4 views of the contact.

<table>
<thead>
<tr>
<th># Reg.</th>
<th>Segmentation IoU CHI3D</th>
<th>Signature IoU CHI3D</th>
<th>Segmentation IoU FlickrCI3D</th>
<th>Signature IoU FlickrCI3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>0.692</td>
<td>0.456</td>
<td>0.472</td>
<td>0.226</td>
</tr>
<tr>
<td>37</td>
<td>0.790</td>
<td>0.542</td>
<td>0.682</td>
<td>0.370</td>
</tr>
<tr>
<td>17</td>
<td>0.815</td>
<td>0.638</td>
<td>0.721</td>
<td>0.499</td>
</tr>
<tr>
<td>9</td>
<td>0.878</td>
<td>0.745</td>
<td>0.799</td>
<td>0.635</td>
</tr>
</tbody>
</table>

Figure 3: 3d contact segmentations and signatures from FlickrCI3D. For an RGB image, the annotators map facets from one mesh to facets on the other mesh if they are in direct contact. By doing so, they automatically segment the regions in contact (marked in red) and facet-level correspondences (marked by arrows).

Figure 4: (Left) Frequency of body regions involved in a contact (75 regions). Note the left-right symmetry and the high frequency for the arms, shoulders and back regions. (Right) Correspondence frequency counts (17 regions).

Datasets Statistics. As the annotation task is expensive and time-consuming, each interaction in the datasets is labeled by only one annotator. Following the annotation process on FlickrCI3D, we gather a number of 11,216 images and 14,081 valid pairs of people in contact, with 81,233 facet-level correspondences within 138,213 selected regions. This results in an average of 5.77 correspondences per pair of people. For CHI3D we gather 2,524 images and pairs of people, with 10,168 facet-level correspondences within 15,168 selected regions. This results in an average of 4.03 correspondences per pair of people.

In fig. 4 (left) we show a 3d human region heat map based on the frequencies of the regions involved in a contact. Notice that the front side of the hands/arms as well as the back side of a person are the most common body parts involved in contacts. This observation is also confirmed by the work of [41] who give contact region maps for various types of human relationships. In fig. 4 (right) we show a frequency map of the correspondences at region level. Notice the large coverage of annotated contact correspondences.

Given the ambiguities of determining contact correspondences from a single view, we check the consistency between annotators on a small common set of images. Results can be seen in Table 1. We evaluate the consistency considering different levels of granularity when grouping facets into regions. It can be noticed that at the highest level of detail they have lower consistency, but as coarser regions of the body are aggregated, consistency increases. This observation is partly congruent to the computational perception study of [26] who argue that humans are not very precise in re-enacting 3d body poses viewed in monocular images. Note that, for CHI3D, consistency is higher as mul-
Figure 5: Multi-task architecture ISP for detection and prediction of interaction signatures, that (1) classifies whether people are in contact, (2) holistically segments the corresponding 3d body surface contact regions for each person, and (3) determines their specific 3d body contact signature. Each task has a specific loss, $L_B$, $L_S$ and $L_C$ respectively. As input, we feed an RGB image of people in close proximity, two 2d skeleton predictions and semantic human features computed on the image. The binary contact estimation task uses a single fully connected layer. The 3d contact segmentation and signature prediction tasks use a sequence of fully connected layers and graph convolutions (shared by both tasks), followed by fully connected layers specialized for each task.

### 3. Methodology

In this section, we describe the models we introduce for the following tasks: (1) contact detection (classification), (2) 3d contact segmentation, (3) 3d contact signature prediction and (4) 3d pose and shape reconstruction using contact signatures.

For tasks (1)-(3), we propose learning methods (collectively referred as Interaction Signature Prediction - ISP) based on deep neural networks that take as input an image $I$ cropped around the bounding box of the two interacting people $P_1$, $P_2$, together with the associated 2d human body poses detected [5]. We encode each 2d body pose as $n_{kp}$ channels, one for each keypoint type, by considering a 2d Gaussian around the coordinate of each keypoint. Following [8], we stack the two pose encodings with the RGB image $I$. In addition, we also stack semantic human features to the input, i.e. 2d body part labeling [34] and 2d part affinity fields, both computed on $I$.

We use the ResNet50 [12] backbone architecture (up to the last average pooling layer) as a trainable feature encoder, which we modify to accommodate the larger number of input channels by increasing the size of the first convolutional filters. An overview of the pipeline is given in fig. 5.

### 3.1. Contact Classification

Given an image $I$ with two people in close proximity we want to estimate if there is any physical contact between the two. We train a deep binary classification network composed of the feature encoder network and add a fully connected layer which outputs the probability of the two label classes, $B = \{0, 1\}$, $1$ – "contact" and $0$ – "no contact". We train using the weighted binary cross entropy loss function with $w_0 < w_1$ as the weights for classes $0$ and $1$, respectively, to account for the more frequent "no-contact" class.

### 3.2. Contact Segmentation and Signature

In order to operate within a manageable output space, we consider contact segmentation and signature prediction at a region-level. In the following, let $N = N_{reg}$, $S = S_{reg} \in \mathbb{R}^{N \times 2}$, the ground-truth contact segmentation, and $C = C_{reg} \in \mathbb{R}^{N \times N}$, the ground-truth contact signature. We leverage the synergy between the segmentation and signature tasks and train them together in a multi-task setting.

Following the feature encoder backbone, we split the network into separate computational pathways for each person, in order to better disentangle their feature representations. As a first step, we extract two sets of features $F_p \in \mathbb{R}^{N \times D_0}$, $p = 1, 2$, using fully connected layers. To integrate the topology of the regions on the mesh, we next apply a fixed number of graph convolution iterations, following the architecture proposed in [17]. The adjacency
matrix we use corresponds to the $N$ regions on the 3d template body mesh, where we set an edge between two regions if they share a boundary. We denote by $F_p$ the output of the graph convolutions. We pass these features to segmentation ($\Theta_S$) and signature ($\Theta_C$) specialization layers, each implemented as a fully connected layer.

The output of the $\Theta_S$ layers, $\bar{S} = [\bar{S}_1 \bar{S}_2]$, represents the final segmentation prediction for the two persons. We use the sigmoid cross-entropy loss

$$L_S(I) = -\sum_{i=1}^{2N} (p_S S_i \log(\bar{S}_i) + (1 - S_i) \log(1 - \bar{S}_i))$$

with a balancing term $p_S \in \mathbb{R}$ between the positive and negative classes. For the contact signature prediction task, we use the output of the $\Theta_C$ layers, $F'_p$, and compute our estimate as $\bar{C} = F'_1 \ast F'_2^T$. We again use the cross entropy loss, $L_C$, with the difference that we iterate to $N \times N$ and use another balancing term, $p_C \in \mathbb{R}$.

### 3.3. 3D Reconstruction with Contact Signatures

Given an image $I$ and its contact signature $C(P_1, P_2)$, we want to recover the 3d pose and shape parameters of the two people in contact, $P_1$ and $P_2$. We start from the optimization framework of [50] and augment it with new loss terms that explicitly use the contact signature. The original energy formulation used in [50] is given by

$$L = \sum_{i \in \{1, 2\}} (L_S(P_1) + L_{psr}(P_1)) + L_{col}(P_1, P_2)$$

where $L_S$ corresponds to the 2d semantic projection error with respect to the visual evidence extracted from the image (i.e. semantic body part labeling and 2d pose) and $L_{psr}$ is a term for pose and shape regularization. $L_{col}(P_1, P_2)$ is a 3d collision penalty between $P_1$ and $P_2$ computed on a set of bounding sphere primitives. Gradients are passed from the loss function, through the 3d body model, all the way to the pose and shape parameters. Our body model has articulated body and hands [47, 35] and we additionally use estimated 2d hand joints positions. This is necessary when modeling two interacting people, as hands are often involved in physical contact (see fig. 4). We reflect these changes in the adapted terms $L_{'S}$ and $L_{psr}^*$, and also introduce a new contact signature loss term, $L_C(P_1, P_2)$, that measures the geometric alignment of regions in correspondence. The adapted energy formulation becomes

$$L^* = \sum_{i \in \{1, 2\}} (L_{'S}(P_1) + L_{psr}^*(P_1)) + L_{col}(P_1, P_2) + L_C(P_1, P_2)$$

where $L_C(P_1, P_2) = L_D(P_1, P_2) + L_N(P_1, P_2)$. The first term $L_D(P_1, P_2)$ seeks to minimize the sum of the distances between all the region pairs that are in contact, $(r_1, r_2) \in C(P_1, P_2)$.

$$L_D(P_1, P_2) = \sum_{(r_1, r_2) \in C(P_1, P_2)} \Phi_D(r_1, r_2)$$

where the distance between two regions $\Phi_D(r_1, r_2)$ is

$$\Phi_D(r_1, r_2) = \sum_{f_1 \in \Psi_D(r_1)} \min_{f_2 \in \Psi_D(r_2)} \phi_D(f_1, f_2)$$

For a facet in one region, this function takes its respective first nearest neighbor facet in the second region and computes the Euclidean distance $\phi_D(f_1, f_2)$ between the centers of the two facets, $f_1$ and $f_2$. Our approach is similar to iterative closest point, but performed at facet level. For each region, we consider a subset of facets obtained by applying a selection operator $\Psi_D$. This offers flexibility to operate not only on the entire set of facets (computationally intensive), but also on a fixed number of uniformly sampled facets or on a given subset of facets, e.g. in the case of ground-truth facet level correspondences.

The second term, $L_N(P_1, P_2)$, enforces the orientation alignment for all region surfaces in contact, $(r_1, r_2) \in C(P_1, P_2)$.

$$L_N(P_1, P_2) = \sum_{(r_1, r_2) \in C(P_1, P_2)} \Phi_N(r_1, r_2)$$

Table 2: Results of our contact segmentation and signature estimation on FlickrCI3D, evaluated for different region granularities on the human 3d surface (from 75, down to 9 regions). We ablate different components of our full method to illustrate their contribution. Human performance represents the consistency values between annotators from Table 1.
Table 3: 3d human pose and translation estimation errors, as well as mean (median) 3D contact distance, expressed in mm, for the CHI3D dataset. We report quantitative results on our collected FlickrCI3D dataset. We split both the contact classification database (90, 167 images) and the contact segmentation and correspondences database (14, 081 images) into train, validation and test subsets each, using the following proportions 85%, 7.5% and 7.5% respectively. In all our experiments, we validate the meta-parameters on the validation set and report the results on the test set. We evaluate the performance of the contact detection task and obtain an average accuracy of 0.846, with 0.844 for the “contact” class and 0.848 for the “no contact” class.

For the contact segmentation and signature prediction method, we train our network with $N_{reg} = 75$, though we can also obtain the coarser versions post-hoc. In training, since our ground truth does not necessarily contain the full set of region correspondences, we do not penalize the non-annotated (but possible) correspondences between the segments on one person and those on the other. At inference time, we exploit the contact segmentation and use its predictions to mask spurious correspondences.

We evaluate our predictions using the intersection over union (IoU$_{N_{reg}}$) metric, computed for different region granularities. Table 2 reports the performance of our full model, for which predictions get closer to the human performance as the region granularity becomes coarser. We also train a version of our method without concatenating the semantic 2d features to the input. In almost all cases, these input features affect performance positively. Similarly, jointly learning the two tasks and using the contact segmentation mask to eliminate non-valid correspondences improves the contact signature estimation performance.

**3D Reconstruction Results.** In fig. 6 we show reconstruction examples on images from FlickrCI3D using our proposed model formulation (see (3)) and contact annotations at different levels of granularity. The first column represents an RGB image from the FlickrCI3D dataset. Without contact information, the method of [50] may provide plausible poses, sometimes with good image alignment, but the subtleties of the interaction are lost. In order to avoid mutual volume intersections, people may even be placed far from each other. By incorporating contact correspondences we can successfully recover the essential details of the interactions, such as hands touching during a handshake/dancing or tackling someone in a rugby match.

We also perform quantitative experiments on the CHI3D dataset and ablate the impact of our geometric alignment term, $L_G$. We select a small set of images (160) to validate the weight of each term in the energy function and report results on the remaining test set. We evaluate both the inferred pose, using mean per joint position error (MPJPE), and the estimated translation. In addition, on the estimated reconstructions, we compute the mean and median distance between the regions in the contact signature. This contact distance is defined as the minimum Euclidean distance between each pair of facets from two regions annotated to be in correspondence. Results are given in table 3 where
Figure 6: 3D human pose and shape reconstructions using contact constraints of different granularity. The first column shows the RGB images followed by their reconstructions without contact information (column 2), using contacts based on 37 and 75 regions, respectively (columns 3 & 4), and using facet-based correspondences (column 5). While using facet-based constraints provides the most accurate estimates, reasonable results can be obtained even for coarser (region) assignments.

Annotated contact information improves the accuracy of the reconstruction. For pose, we only evaluate on a standard 3D body joints configuration compatible with the MoCap format that does not include body extremities (e.g., hands and feet). Our complete optimization framework not only produces more accurate reconstructions of pose and translation, but also closely approaches the contact signature.

5. Conclusions

We have argued that progress in human sensing and scene understanding would eventually require the detailed 3D reconstruction of human interactions where contact plays a major role, not only for veridical estimates, but in order to ultimately understand fine-grained actions, behavior and intent. We have proposed a graded modeling framework for Interaction Signature Prediction (ISP) based on contact detection and 3D correspondence estimation over model surface regions at different levels of detail, with subsequent 3D reconstruction under losses that integrate contact and surface normal alignment constraints. We have undertaken a major effort to collect 3D ground truth data of humans involved in interactions (CHI3D, 631 sequences containing 2,525 contact events, 728,664 ground truth poses), as well as image annotations in the wild (FlickrCI3D, a dataset of 11,216 images, with 14,081 processed pairs of people, and 81,233 facet-level surface correspondences within 138,213 selected regions). We have evaluated all components in detail, showing their relevance towards accurate 3D reconstruction of human contact. Models and data are made available for research.

Acknowledgments: This work was supported in part by the ERC Consolidator grant SEED, CNCS-UEFISCDI (PN-III-P4-ID-PCCF2016-0180) and SSF.
References

[1] CMU graphics lab. CMU graphics lab motion capture database. 2009. http://mocap.cs.cmu.edu/.


Leonid Sigal, Alexandru O Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV, 2010.


