Deep Multitask Architecture for Integrated 2D and 3D Human Sensing

Alin-Ionut Popa[2]* Mihai Zanfir[2]* Cristian Sminchisescu[1,2]
alin.popa@imar.ro, mihai.zanfir@imar.ro cristian.sminchisescu@math.lth.se
[1]Department of Mathematics, Faculty of Engineering, Lund University
[2]Institute of Mathematics of the Romanian Academy

Abstract

We propose a deep multitask architecture for fully automatic 2d and 3d human sensing *(DMHS), including* recognition and reconstruction*, in* monocular images*. The system computes the figure-ground segmentation, semantically identifies the human body parts at pixel level, and estimates the 2d and 3d pose of the person. The model supports the joint training of all components by means of multi-task losses where early processing stages recursively feed into advanced ones for increasingly complex calculations, accuracy and robustness. The design allows us to tie a complete training protocol, by taking advantage of multiple datasets that would otherwise restrictively cover only some of the model components: complex 2d image data with no body part labeling and without associated 3d ground truth, or complex 3d data with limited 2d background variability. In detailed experiments based on several challenging 2d and 3d datasets (LSP, HumanEva, Human3.6M), we evaluate the sub-structures of the model, the effect of various types of training data in the multitask loss, and demonstrate that state-of-the-art results can be achieved at all processing levels. We also show that in the wild our monocular RGB architecture is perceptually competitive to a state-of-the art (commercial) Kinect system based on RGB-D data.*

1. Introduction

The visual analysis of humans has applications as diverse as autonomous vehicles, robotics, human-computer interaction, virtual reality, and digital libraries, among others. The problem is challenging due to the large variety of human poses and body proportions, occlusion, and the diversity of scenes, angles of observation, and backgrounds humans are pictured against. The *monocular* case, which is intrinsic to many scenarios like the analysis of photographs or video available on the web, adds complexity as depth information is missing for 3d reconstruction. This leads to geometric ambiguity and occlusion which are difficult to resolve compared to situations where multiple cameras are present.

A detailed analysis at both 2d and 3d levels, further exposes the need for both measurement and prior-knowledge, and the necessary inter-play between segmentation, reconstruction, and recognition within models that can jointly perform all tasks. This is one of our objectives.

As training is essential, a major difficulty is also the limited coverage of current datasets: 2d repositories like LSP [18] or MPI-II [3] exhibit challenging backgrounds, human body proportions, clothing, and poses, but offer single viewpoints, provide only approximate 2d joint location ground truth, and do not carry human segmentation or body part labeling information. Their size is also relatively small by today's deep learning standards. In contrast, 3d datasets like HumanEva [39] or Human3.6M [16] offer extremely accurate 2d and 3d anatomical joint or body surface reconstructions and a variety of poses captured under multiple viewpoints. Human 3.6M is also large-scale. However, being captured indoors, the 3d datasets typically lack the background and clothing variability that represent a strength of the 2d datasets captured in the wild, although the situation is slightly more nuanced. Some of the 3d datasets (e.g. Human3.6M) come with mixed-reality training setups where a moderately realistic graphics character is placed, with a geometrically correct setup, into a real scene and animated using human motion capture. Arguably, though, a complete fully realistic training setting that contains accurate 2d and 3d information is still elusive. An open question is how one can leverage the separate strengths of existing 2d and 3d datasets towards training models that can operate in challenging images and offer accurate recognition and reconstruction estimates.

In this paper we propose one such deep learning model which, given a monocular RGB image, is able to fully automatically sense the humans at multiple levels of detail: figure-ground segmentation, body-part labeling at pixel level, as well as 2d and 3d pose estimation. By designing multi-task loss functions at different, recursive processing stages (human body joint detection and 2d pose estima-

*Authors contributed equally

tion, semantic body part segmentation, 3d reconstruction) we are able to tie complete, realistic training scenarios by taking advantage of multiple datasets that would otherwise restrictively cover only some of the model component training (complex 2d image data with no body part labeling and without associated 3d ground truth, or complex 3d data with limited 2d background variability), leading to covariate shift and a lack of model expressiveness. In extensive experiments, including ablation studies performed using representative 2d and 3d datasets like LSP, HumanEva, or Human3.6M, we illustrate the model and show that state-of-the-art results can be achieved for both semantic body part segmentation and for 3d pose estimation.

2. Related Work

This work relates to 2d and 3d monocular human pose estimation methods as well as semantic segmentation using fully-trainable deep processing architectures. As prior-work is comprehensive in each sub-domain [36], we will here mostly cover some of the key techniques directly relating to our approach, with an emphasis towards deep architectures and methodologies aiming to integrate the different levels of 2d and 3d processing.

The problem of 2d human pose estimation has been approached initially using the pictorial structures and deformable part models where the kinematic tree structure of the human body offers a natural decomposition [11, 19, 7, 49, 29, 12]. In recent years, deep learning had a great impact over the state-of-the-art pose estimation models where different hierarchical feature extraction architectures have been combined with spatial constraints between the human body parts [46, 10, 45, 20, 30, 28, 23, 8, 13, 9, 43, 27, 24]. Recent deep architectures are obtained by cascading processing stages with similar structure but different parameters, where the system combines outputs of early layers with new information extracted directly from the image using learnt feature extractors. Such recursive schemes for 2d pose estimation appear in the work of [33, 48, 9] whereas similar earlier ideas of feeding-back 3d estimates into 2d inference layers appears earlier in [15, 41]. Note that the iterative processing frameworks of [33, 15] are not deep, built on parts-based graphical models and random forests, respectively.

There is a vast literature on monocular 3d human pose estimation, including the analysis of kinematic ambiguities associated with the 3d pose [42, 5], as well as generative and discriminative methods for learning and inference [35, 37, 4, 1, 26, 38, 14, 30]. More recently, deep convolutional architectures have been employed in order to estimate 3d pose directly from images[21, 22, 34, 51] mostly in connection with 3d human motion capture datasets like HumanEva, Human3.6M, where the poses are challenging but the backgrounds are relatively simple. There is also interest in combining 2d and 3d estimation methods in order to obtain models capable of multiple task, and able to operate in realistic imaging conditions [41, 32, 47, 2]. The most recent methods, [6, 34], rely on an a-priori 3d model that is fitted to anatomical body joint data obtained from an initial 2d pose estimate produced by a deep processing architecture, like the one of [13] or [45]. The methods rely on a state-of-the-art discriminative person detection and 2d pose estimation and introduce a generative fitting component to search the space of admissible body proportion variations. Both approaches fit a statistical body shape and kinematic model to data, using one or multiple views and the joint assignment constraints from 2d human pose estimation. These methods use the 3d to 2d anatomical landmark assignment provided by [6, 34] as initialization for 3d inference. While this is effective, as shown in several challenging evaluation scenarios, the initial 3d shape and kinematic configuration still needs to be initialized manually or set to a key initial pose. This is in principle still prone to local optima as the monocular 3d human pose estimation cost is non-linear and non-convex even under perfect 3d to 2d model-to-image assignment [42].

We share with [41, 15, 6, 34] the interest in building models that integrate 2d and 3d reasoning. We propose a fully trainable discriminative model for human recognition and reconstruction at 2d and 3d levels. We do not estimate human body shape, but we do estimate figure-ground segmentation, the semantic segmentation of the human body parts, as well as the 2d and 3d pose. The system is trainable, end-to-end, by means of multitask losses that can leverage the complementary properties of existing 2d and 3d human datasets. The model is fully automatic in the sense that both the human detection and body part segmentation and the 2d and 3d estimates are the result of recurrent stages of processing in a homogeneous, easy to understand and computationally efficient architecture. The approach is complementary to [6, 34]: our model can benefit from a final optimization-based refinement and it would be useful to estimate the human body shape. In contrast, [6, 34] can benefit from the semantic segmentation of the human body parts for their shape fitting, and could use the accurate fully automatic 2d and 3d pose estimates we produce as initialization for their 3d to 2d refinement process.

3. Methodology

In this section we present our multitask multistage architecture. The idea of using multiple stages of recurrent feed-forward processing is inspired by architectures like [33, 48] which focus separately on the 2d and 3d domains. However, we propose an uniform architecture for joint 2d and 3d processing that no prior method covers. Our choice of multi-task loss also makes it possible to exploit the complementary advantages of different datasets.

Conceptually, each of our stages of processing produces recognition and reconstruction estimates and is constrained by specific training losses. Specifically, each stage t is split into semantic processing S^t, and 3d reconstruction, R^t (see fig. 1). At the same time, the semantic module S^t is divided in two sub-tasks, one focusing on 2d pose estimation, J^t, and the other on body part labeling and figure-ground segmentation, B^t (see fig. 2). The first (*i.e.* J^t) feeds into the second one (*i.e.* B^t), while the semantic stages feed into the reconstruction stages. Each task consists of a total of six recurrent stages which take as input the image, the results of previous stages of the same type (except for the first one), as well as inputs from other stages (2d pose estimation feeding into semantic body part segmentation, and both feeding into 3d pose reconstruction). The inputs to each stage are individually processed and fused via convolutional networks in order to produce the corresponding outputs.

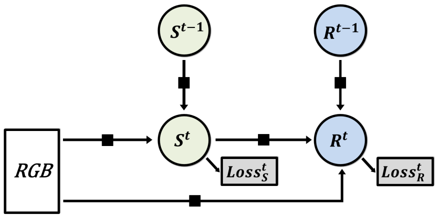

Figure 1. Stage t of our architecture for recognition and reconstruction: figure-ground segmentation, 2d pose estimation, and semantic segmentation of body parts, all denoted by (S), and 3d reconstruction (R). The semantic task is detailed in fig. 2 and fig. 3; the 3d reconstruction task is detailed in fig. 4.

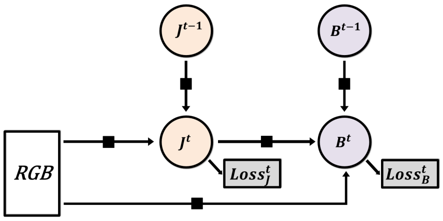

Figure 2. Stage t of our semantic task, including 2d joint detection (J), and labeling of the body parts (B).

3.1. 2D Human Body Joint Detection

The 2d pose estimation task is based on a recurrent convolutional architecture similar to [48]. Given an RGB image $I \in \mathbb{R}^{w \times h \times 3}$, we seek to correctly predict the locations of N_J anatomically defined human body joints $p_k \in \mathcal{Z} \subset \mathbb{R}^2$, with $k \in \{1 \dots N_J\}$. At each stage $t \in \{1 \dots T\}$, where T is the total number of stages, the network outputs belief maps $J^t \in \mathbb{R}^{w \times h \times N_J}$. The first stage of processing operates only on image evidence (a set of seven convolution and three pooling layers producing features $\mathbf{x}$) but for subsequent stages, the network also considers the information in the belief maps fed from the previous stage J^{t-1} with a slightly different image feature function $\mathbf{x}'$, defined as a set of four convolution and three pooling layers as in [48]. These features are transformed through a classification function c_J^t to predict the body joint belief maps J^t. The function c_J^t consists of a series of five convolutional layers, the first three of the form $(11 \times 11 \times 128)$, followed by a $(1 \times 1 \times 128)$ convolution and a final $(1 \times 1 \times N_J)$ convolution that outputs J^t. The loss function at each stage L_J^t minimizes the squared Euclidean distance between the predicted and ground truth belief maps, J^t and $J^\star$:

$$L_J^t = \sum_{k=1}^{N_J} \sum_{z \in \mathcal{Z}} \left\| J^t(z,k) - J^\star(z,k) \right\|_2^2 \qquad (1)$$

In practice, this component of the model can be trained with data from both 2d and 3d datasets like LSP, where the ground truth is obtained manually, or HumanEva and Human3.6M where the ground truth is obtained automatically based on anatomical markers.

3.2. Semantic Body Part Segmentation

In semantic body part segmentation (body part labeling) we assign each image location $(u,v) \in \mathcal{Z} \subset \mathbb{R}^2$ one of N_B anatomical body part labels (including an additional label for background), b_l, where $l \in \{1 \dots N_B\}$. At each stage t, the network predicts, for each pixel location, the presence probability of each body part, $B^t \in \mathbb{R}^{w \times h \times N_B}$. Differently from the previous task, our aim now is to classify each pixel location, not only to identify the body joints. The loss function used changes from a squared Euclidean to a multinomial logistic:

$$L_B^t = -\frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} \log(B^t_{z,B^\star_z}) \qquad (2)$$

where $B_z^\star$ is the ground truth label for each image location $z = (u,v)$.

During the first stage of processing, we use convolutional representations based on the image (a series of convolution and pooling layers $\mathbf{x}$ with parameters tied from §3.1) and the 2d pose belief maps J^1 in order to predict the current body labels B^1. For each of the following stages, we also use the information present in the body labels at the previous stage, B^{t-1}, and rely on a series of four convolutional layers c_B^t that learn to combine inputs obtained by stacking image features $\mathbf{x}$ and B^{t-1}. The function c_B^t shares the same structure as the first four convolutions in c_J^t, but a classifier in the form of a $(1 \times 1 \times N_B)$ convolution is applied after the fusion with the current 2d pose belief maps J^t, in order to obtain semantic probability maps B^t. An overview of our architecture together with the main dependencies is given in figure 3. Finally, we use an additional deconvolution layer [25] of size $16 \times 16 \times N_B$, such that the loss can be computed at the full resolution of the input image I.

In practice, realistic data for training this component of the loss is not as easy to obtain as 2d body joint positions. Human3.6M offers such training data, but we are also able to generate it automatically (approximately) for LSP (§4).

3.3. 3D Pose Reconstruction

This model component is designed for the stage-wise, recurrent reconstruction of 3d human body configurations represented as set of N_R 3d skeleton joints, from a single monocular image I. The estimate is obtained from the internal representations R^t. The 3d reconstruction module leverages information provided by the 2d semantic components S^t, incorporating the joint and body part labeling feature maps J^t and B^t. Additionally, we insert a trainable function c_D^t, defined similarly to c_B^t, over image features, in order to obtain body reconstruction feature maps D^t. The module follows a similar flow as the previous ones: it reuses estimates at earlier processing stages, R^{t-1}, together with S^t and D^t, in order to predict the reconstruction feature maps R^t. The processing stages and dependencies of this module are shown in fig. 4.

Procedurally, we first fuse S^t and D^t, then apply a series of single $(3 \times 3 \times 128)$, $(3 \times 3 \times 64)$, $(1 \times 1 \times 64)$ convolutions, followed by a pooling layer (3×3) and a $(1 \times 1 \times 16)$ convolution. The output is concatenated with R^{t-1} and convolved by a $(1 \times 1 \times 16)$ kernel that learns to combine the two components, producing the estimate R^t. The feature maps are then transformed to the desired dimensionality of the 3d human body skeleton by means of a fully connected layer. The loss L_R^t is expressed as the mean per joint position error (MPJPE):

$$L_R^t = \sum_{i=1}^{N_R} \sqrt{\sum_{j=1}^{3} (f(R^t, i, j) - R^\star(i, j))^2 + \epsilon^2} \qquad (3)$$

where $R^\star$ are the 3d ground truth human joint positions, $f(\cdot)$ is the fully connected layer applied to R^t and ϵ is a small constant that makes the loss differentiable.

This loss component can be trained with data from HumanEva and Human3.6M, but not from LSP or other 2d datasets as these lack 3d ground truth information. Although the backgrounds in HumanEva and Human3.6M are not as challenging as those in LSP, the use of a multitask loss makes the complete 2d and 3d model competitive not only in the laboratory but also in the wild (§4 and fig. 5).

3.4. Integrated Multi-task Multi-stage Loss

Given the information provided in the previous sections, we are now able to define the complete multitask, multi-stage loss function of the model as follows:

$$L = \sum_{t=1}^{T} (L_J^t + L_B^t + L_R^t) \qquad (4)$$

The loss allows us to conveniently train all the model component parameters, for different tasks, based on datasets where the annotations are challenging, or where annotations are missing entirely, as datasets with different coverage contribute to various components of the loss. Whenever we train using datasets with partial coverage, we could freeze the model components for which we do not have ground-truth. We can also simultaneously train all parameters, using datasets with partial and complete coverage: those examples for which we have ground truth at all levels will contribute to each of the loss components, whereas examples for which we have partial ground truth will only contribute to their corresponding losses.

4. Experiments

In order to evaluate our method, we use 3 well-known datasets, the Leeds Sports Dataset (LSP) [18], HumanEva [39] and Human3.6M [16].

The LSP dataset consists of 2d pose annotated RGB images depicting people doing sports (*athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, volleyball*). We use both the original release containing $1,000$ training and $1,000$ testing images, as well as the extended training release containing an additional $10,000$ images.

We also use the HumanEva-I dataset, obtained by an accurate 3d motion capture system in the laboratory. There are six actions performed in total by three subjects. As standard procedure [40, 4, 50, 44], we train our model on the train set and report results on the validation set, where we only consider every 5^{th} frame of the sequences *walking*, *jog* and *box* for all three subjects and a single frontal camera view.

Human80K is an 80,000 sample subset of the much larger 3.6 million human pose dataset Human3.6M [16]. The dataset is captured in a laboratory environment with a motion capture setup, and contains daily activities and interaction scenarios (*providing directions, discussion, eating, activities while seating, greeting, taking photo, posing, making purchases, smoking, waiting, walking, sitting on chair, talking on the phone, walking dog, walking together*). The actions are performed by 11 actors, and captured by 4 RGB cameras. The dataset is fully annotated and contains RGB data, 2d body part labeling ground truth masks as well as accurate 2d and 3d pose reconstructions. Human80K consists of $55,144$ training and $24,416$ testing samples from Human3.6M. The samples from each original capture were selected such that the distance between each-other, in 3d space, is no more than 100 mm.

We implement our models in Caffe [17]. The complete recognition and reconstruction pipeline takes approxi-

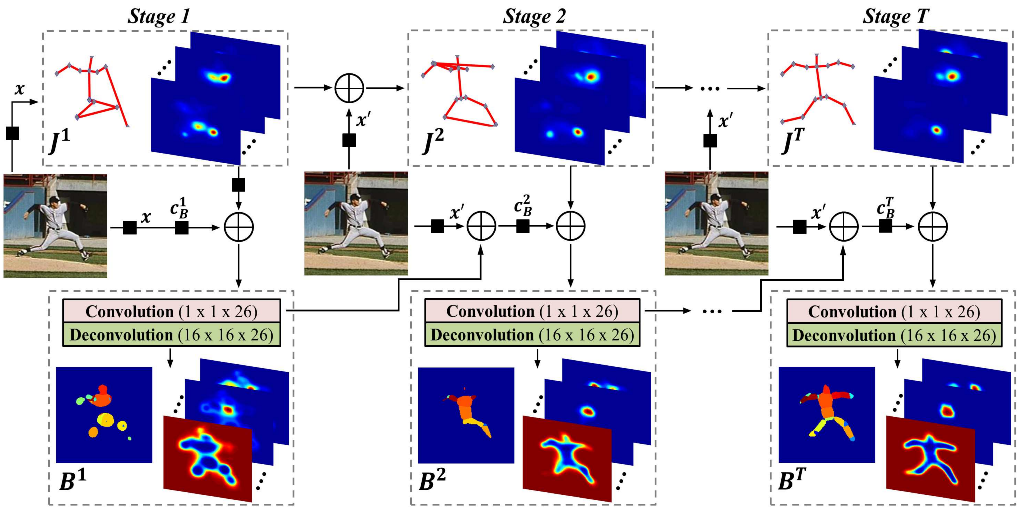

Figure 3. Our multitask multistage 2d semantic module S^t, combines semantic body part labeling B^t and 2d pose estimation J^t.

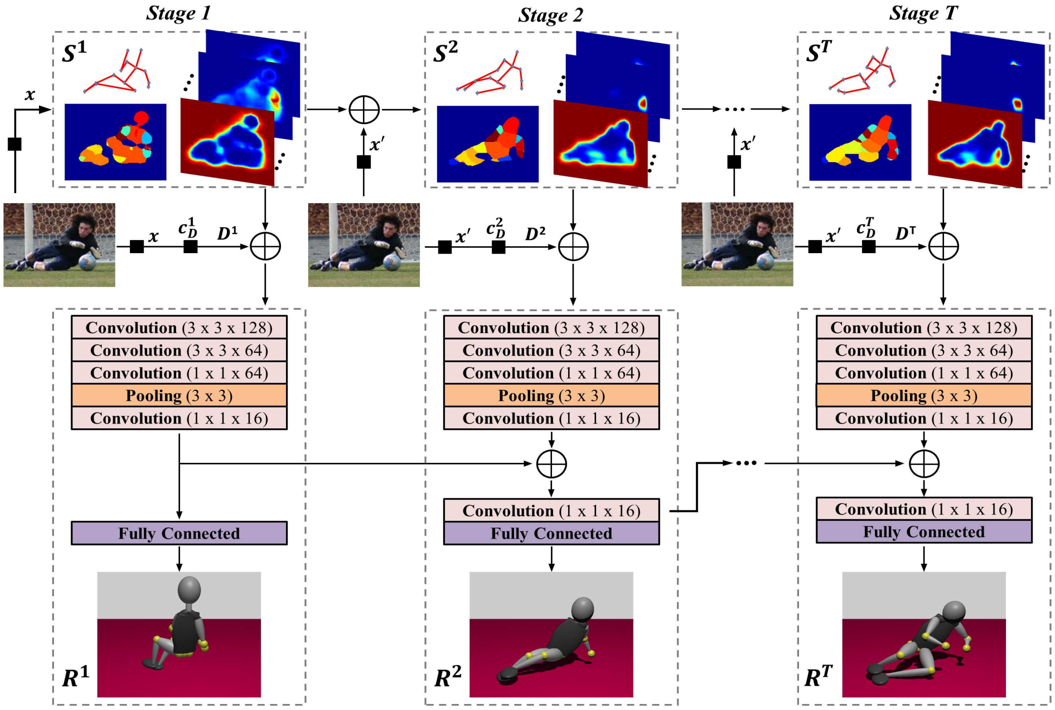

Figure 4. Our multitask multistage 3d reconstruction module R^t, combines 3d processing with information from semantic modules, S^t.

mately 400 ms per frame, in testing, on an Nvidia TITAN X (Pascal) 12GB GPU. We evaluate both the recognition (2d body part labeling) and the 3d reconstruction capabilities of our architecture. We use $T = 6$ stages in our architecture for each sub-task component model (joint detection, semantic segmentation, 3d reconstruction) and report results only for the final stage of each sub-task, as it is the best performing according to validation.

4.1. Body Part Labeling

In order to evaluate the 2d body part labeling task, we use the Human80K and LSP datasets. We introduce additional annotations for LSP, as they are not available with the original release which only provides 2d anatomical joints. We create human body part annotations for LSP by using the annotated 2d skeleton joints and the kinematic tree. We produce circles for skeleton joints and ellipses for individual body parts. We set the major axis to the size of the segment between the corresponding joints and estimate a constant value, for each body part, for the minor axis. The reason for augmenting LSP, is to enhance the variability in human appearance, body proportions and backgrounds in addition to those available in H3.6M.

For evaluation on Human80K, we compare with the results of [15] which represent the state-of-the-art for this task on this dataset. The authors of [15] assume that the silhouette of the person (the figure-ground segmentation) is given, and perform body part labeling only on the foreground mask as an inference problem over a total of 24 labels. Differently from them, we do not make this assumption and consider the background as an extra class, thus building a model that predicts 25 classes.

To extend the evaluation on LSP, we consider multiple scenarios: (a) training on Human80K with (b) fine-tuning on LSP and (c) training our architecture on LSP and Human80K simultaneously and test on both Human80K and LSP. In our setup, training using only LSP was not feasible, as in multiple experiments the network did not converge. The labeling parameters (stages B) are initialized randomly, while the parameters corresponding to the 2d joint detection components, J, are initialized with the values of the network presented in [48], trained on MPI-II and LSP.

The performance of our body part labeling models for Human80K and LSP is given in tables 1 and 2. We use the same evaluation metrics as in [15], *i.e.* average accuracy for the pixels contained within the ground truth silhouette and class normalized average accuracy. As these two metrics apply only to the foreground classes, we also compute for all pixels, background and foreground, the average accuracy as well as the class normalized average accuracy.

From table 1 (Human80K testing), it can be noted that even though we solve a harder problem (by additionally estimating the figure-ground segmentation), the class normalized average precision greatly improves, with more than 10% over [15] for the models trained on (a) Human80K and (c) Human80K and LSP jointly. However, the model (b) initialized with parameters trained on Human80K, but fine-tuned on LSP, seems to have a performance drop caused by the low quality of the LSP body label annotations. In this scenario, as expected, the best performance is obtained by the model trained on Human80K, perhaps due to the fact that the test set distribution is better captured by the unaltered Human80K training set. This is not the case when testing on LSP, as it can be seen from table 2. The best performance is obtained by the model trained jointly on Human80K and LSP. This model is able to combine the pose variability and part labeling annotation quality of Human80K with the background and appearance variability of LSP, making it adequate for both laboratory settings and for images of people captured against challenging backgrounds. [1]

The proposed models for the 2d body part labeling task are trained using an initial learning rate of 10^{-10}, reduced every 5 epochs by a factor $\gamma = 0.33$. During the learning process, the training data is augmented by randomly rotating with angles between $[-40°, +40°]$, scaling by a factor in the range $[0.5, 1.2]$ and horizontally flipping, in order to improve the diversity of the training set. Qualitative results of the semantic body part segmentation in challenging images are shown in fig. 5.

4.2. 3D Human Pose Reconstruction

For the evaluation of 3d pose reconstruction we use HumanEva-I and Human3.6M, as they are both challenging datasets, containing complementary poses, and offering accurate 3d pose annotations. In all experiments, we use the same 3d evaluation metrics as other methods we compare against, according to standard practice. For the following experiments, note that our 2d semantic module components are trained on data from LSP and Human80K, whereas the 3d component of the module was pre-trained using Human80K. For our 3d network, we use an initial learning rate set to 10^{-7} and reduce it at every 5 epochs by a constant factor $\gamma = 0.66$.

We report results on the Human80K test set and study the impact of each of the input features J, B and D on the 3d reconstruction task (see table 4). Notice that models that use the 'D-pipeline' for the 3d reconstruction part, would correspond to convolutional networks that feed-forward from the image features, but do not leverage the semantic segmentation of human body parts.

We observe that our fully integrated system, DMHS_R(J,B,D), achieves the lowest error of 63.35 mm. Besides the interest in computing additional detailed semantic human representations, it can be seen that by feeding in the results from other tasks – the body joint belief maps J and the body labeling probability maps B – the error is reduced considerably, from 77.56 mm (for a model based on feed-forward processing and infering additional 2d joint position information) to 63.35 mm. Notice also our significant gain with respect to previous state-of-the-art results on Human80K, as reported by [15].

For HumanEva-I, we use the DMHS_R(J,B,D) architecture trained on Human80K and fine-tuned for a few epochs (6) on the HumanEva-I training data. We use a subset of the training set containing $4,637$ samples. The fine-tuning step on HumanEva-I is performed in order to compensate for the differences in marker positioning with respect to Human80K, and in order to account for the different pose distributions w.r.t. Human80K. In table 3, we compare against

[1] When training and testing distributions do not shift (e.g. Human80K) additional data from a different distribution doesn't necessarily help. But additional data collected in the lab helps in the wild.

Avg. Acc. (%)	DMHS$_{\text{B}}$ - Human80K	DMHS$_{\text{B}}$ - LSP (ft)	DMHS$_{\text{B}}$ - Human80K & LSP	[15]
Per pixel (fg)	**79.00**	53.31	75.84	73.99
Per pixel (fg + bg)	**91.15**	83.30	89.92	-
Per class (fg)	**67.35**	43.40	64.83	-
Per class (fg + bg)	**68.56**	45.61	66.13	53.10

Table 1. **Body part labeling results for the Human80K test set**. We report the performance of our network, trained on Human80K, LSP and both datasets jointly. Note that for LSP, the network was pre-trained on Human80K, as it otherwise failed to converge trained on LSP alone. We also compare with [15], where comparisons are possible only for accuracies computed on the person foreground as the model of [15] does not predict the background label. Our model is able to predict the background class, thus we report the performance for the entire image (including the background class) as well as performance inside the silhouette (for classes associated to human body parts). Note that the best performance on Human80K is obtained with the network trained on Human80K, perhaps due to the noisy annotations added to LSP, and the naturally more similar training and testing distributions of Human80K.

Avg. Acc. (%)	DMHS$_{\text{B}}$ - Human80K	DMHS$_{\text{B}}$ - LSP (ft)	DMHS$_{\text{B}}$ - Human80K & LSP
Per pixel (fg)	50.52	60.54	**61.16**
Per pixel (fg + bg)	85.58	91.08	**91.09**
Per class (fg)	36.46	44.73	**45.91**
Per class (fg + bg)	38.77	46.88	**48.01**

Table 2. **Body part labeling results for the LSP dataset**. All models are initialized on the Human80K dataset, as networks trained only using LSP failed to converge. In this case the models trained jointly on Human80K and LSP produced the best results. This shows the importance of having accurate body part labeling annotations, obtained in simple imaging scenarios but for complex body poses, in combination with less accurate annotations but with more complex foreground and background appearance variations.

Method	Walking			Avg.	Jog			Avg.	Box			Avg.	All
[40]	65.1	48.6	73.5	62.4	74.2	46.6	32.2	51.0	-	-	-	-	-
[4]	45.4	28.3	62.3	45.33	55.1	43.2	37.4	45.2	42.5	64.0	69.3	58.6	49.7
[50]	35.8	32.4	41.6	36.6	46.6	41.4	35.4	41.1	-	-	-	-	-
[44]	37.5	25.1	49.2	37.3	-	-	-	-	50.5	61.7	57.5	56.6	-
DMHS$_{\text{R}}$(J,B,D)	27.1	18.4	39.5	**28.3**	37.6	28.9	27.6	**31.4**	30.5	45.8	48.0	**41.5**	**33.7**

Table 3. **3d mean joint position error on the HumanEva-I dataset**, computed between our predicted joints and the ground truth after an alignment with a rigid transformation. Comparisons with other competing methods show that DMHS achieves state-of-the-art performance.

several state-of-the-art methods. We follow the standard evaluation procedure in [40, 4, 50, 44] and sample data from the validation set for *walking, jogging and boxing* activities. We use a single camera. We obtain considerable performance gains with respect to the previous state-of-the-art methods on HumanEva, even though we only use a small subset of the available training set. We also evaluated on the official test set of Human3.6M, using the model trained only on H80K with no additional parameter validation obtaining an average error of 73 mm.[2] Qualitative results of our method can be seen in figs. 5 and 6.

Model	Avg. MPJPE (mm)
[15]	92.00
DMHS$_{\text{R}}$(J)	128.05
DMHS$_{\text{R}}$(D)	77.56
DMHS$_{\text{R}}$(J,B)	118.68
DMHS$_{\text{R}}$(J,D)	72.00
DMHS$_{\text{R}}$(J,B,D)	**63.35**

Table 4. **3d mean joint position error on the Human80K dataset**. Different components of our model are compared to [15].

5. Conclusions

We have proposed a deep multitask architecture for *fully automatic 2d and 3d human sensing* (DMHS), including *recognition and reconstruction*, based on *monocular images*. Our system estimates the figure-ground segmentation of the human, detects the human body joints, semantically identifies the body parts, and reconstructs the 2d and 3d pose of the person. The design of recurrent multi-task loss functions at multiple stages of processing supports the principled combination of the strengths of different 2d and 3d datasets, without being limited by their different weaknesses. In experiments we perform ablation studies, evaluate the effect of various types of training data in the multitask loss, and demonstrate that state-of-the-art-results can be achieved at all processing levels. We show that, even in the wild, our monocular RGB architecture is perceptually competitive to state-of-the art commercial RGB-D systems.

Acknowledgments: This work was supported in part by the European Research Council Consolidator grant SEED, CNCS-UEFISCDI under JRP-RO-FR-2014-16, the EU Horizon 2020 Grant #688835 (DE-ENIGMA), and SSF.

[2] Please check the leaderboard at `http://vision.imar.ro/human3.6m/ranking.php` (Testset H36M_NOS10).

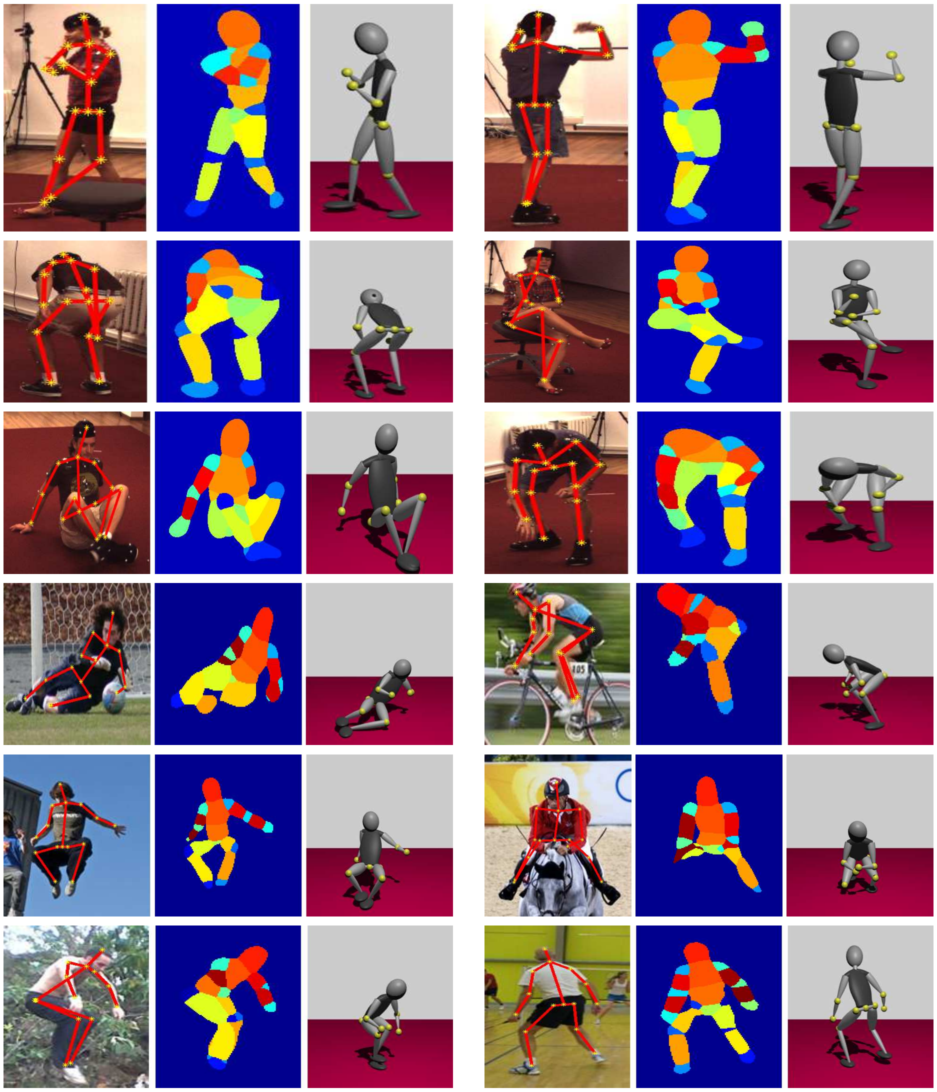

Figure 5. Recognition and reconstruction results for images from Human3.6M and LSP. For each image we show the 2d pose estimate the semantic segmentation of body parts and the 3d pose estimation. Notice the difficulty of backgrounds and poses and the fact that the 2d and 3d models generalize well. Notice that in our architecture, errors during the early stages of processing (2d pose estimation) can be corrected later, during e.g. semantic body part segmentation or 3d pose estimation.

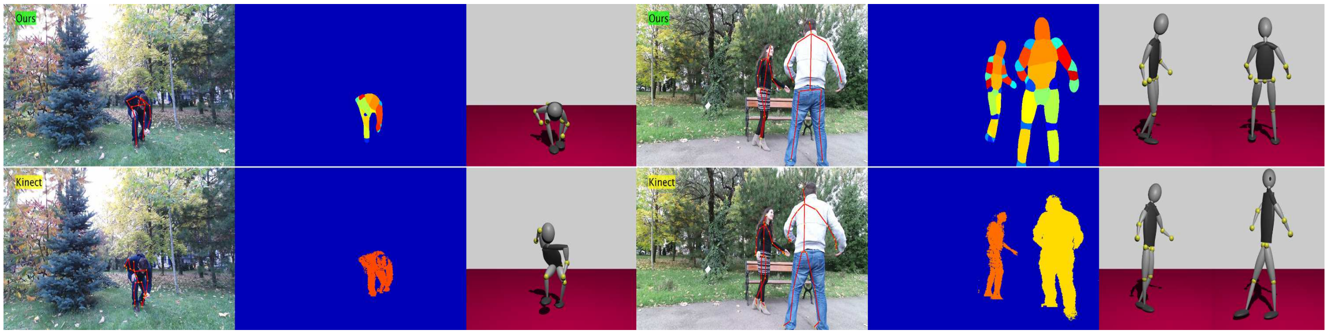

Figure 6. Qualitative comparisons for segmentation and reconstruction between our RGB model (top row) and the ones of a commercial RGB-D Kinect for Xbox One system (bottom row). Our model produces accurate figure-ground segmentations, body part labeling, and 3d reconstruction for some challenging poses.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 28(1):44–58, 2006. 2

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, 2015. 2

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1

[4] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *IJCV*, 87(1-2):28–52, 2010. 2, 4, 7

[5] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas. Fast algorithms for large scale conditional 3d prediction. In *CVPR*, 2008. 2

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2

[7] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 2

[8] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016. 2

[9] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2

[10] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014. 2

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 2

[12] A. Hernández-Vela, S. Sclaroff, and S. Escalera. Poselet-based contextual rescoring for human pose estimation via pictorial structures. *IJCV*, 118:49–64, 2016. 2

[13] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 2

[14] C. Ionescu, L. Bo, and C. Sminchisescu. Structural svm for visual localization and continuous state estimation. In *ICCV*, 2009. 2

[15] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *CVPR*, 2014. 2, 6, 7

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014. 1, 4

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4

[18] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 1, 4

[19] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2

[20] M. Kiefel and P. V. Gehler. Human pose estimation with fields of parts. In *ECCV*, 2014. 2

[21] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 2

[22] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *ICCV*, 2015. 2

[23] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 2

[24] I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. In *ECCV*, 2016. 2

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 4

[26] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *PAMI*, 28(7):1052–1062, 2006. 2

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[28] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 2

[29] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013. 2

[30] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, 2014. 2

[31] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. *arXiv preprint arXiv:1701.08985*, 2017.

[32] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *ECCV*, 2012. 2

[33] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014. 2

[34] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, 2016. 2

[35] R. Rosales and S. Sclaroff. Learning body pose via specialized maps. In *NIPS*, 2001. 2

[36] B. Rosenhahn, R. Klette, and D. Metaxas, editors. *Human Motion, Understanding, Modelling, Capture and Animation*, volume 36. Springer Verlag, 2008. 2

[37] G. Shakhnarovich, P. A. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, 2003. 2

[38] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2007. 2

[39] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 1, 4

[40] 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 *CVPR*, 2013. 4, 7

[41] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning joint top-down and bottom-up processes for 3d visual inference. In *CVPR*, volume 2, pages 1743–1752. IEEE, 2006. 2

[42] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR*, 2003. 2

[43] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicut and deep matching. In *ECCV*, 2016. 2

[44] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. In *CVPR*, June 2016. 4, 7

[45] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 2

[46] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2

[47] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *CVPR*, 2014. 2

[48] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, June 2016. 2, 3, 6

[49] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 35(12):2878–2890, 2013. 2

[50] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *CVPR*, June 2016. 4, 7

[51] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 2016. 2