Single Image 3D Human Pose Estimation from Noisy Observations

E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, F. Moreno-Noguer
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
08028, Barcelona, Spain

Abstract

Markerless 3D human pose detection from a single image is a severely underconstrained problem because different 3D poses can have similar image projections. In order to handle this ambiguity, current approaches rely on prior shape models that can only be correctly adjusted if 2D image features are accurately detected. Unfortunately, although current 2D part detector algorithms have shown promising results, they are not yet accurate enough to guarantee a complete disambiguation of the 3D inferred shape.

In this paper, we introduce a novel approach for estimating 3D human pose even when observations are noisy. We propose a stochastic sampling strategy to propagate the noise from the image plane to the shape space. This provides a set of ambiguous 3D shapes, which are virtually undistinguishable from their image projections. Disambiguation is then achieved by imposing kinematic constraints that guarantee the resulting pose resembles a 3D human shape. We validate the method on a variety of situations in which state-of-the-art 2D detectors yield either inaccurate estimations or partly miss some of the body parts.

1. Introduction

Recovering the 3D human pose from a single image is inherently an ill-posed problem because many different body configurations may have very similar image projections. The problem becomes even more challenging if we consider realistic situations in which image features, such as the body silhouette, limbs or 2D joints, cannot be accurately detected, either due to self occlusions or to the presence of distracting backgrounds. This is the scenario we contemplate, and which we will tackle in a two step process: first we will consider an off-the-shelf detector [37] to estimate the positions of 2D body parts. As shown in Fig. 1 the output of this algorithm is a set of bounding boxes for each body part, whose center may contain relatively large deviations from the true positions. In a second stage, we will propose a methodology to filter out these artifacts and estimate an accurate 3D body pose.

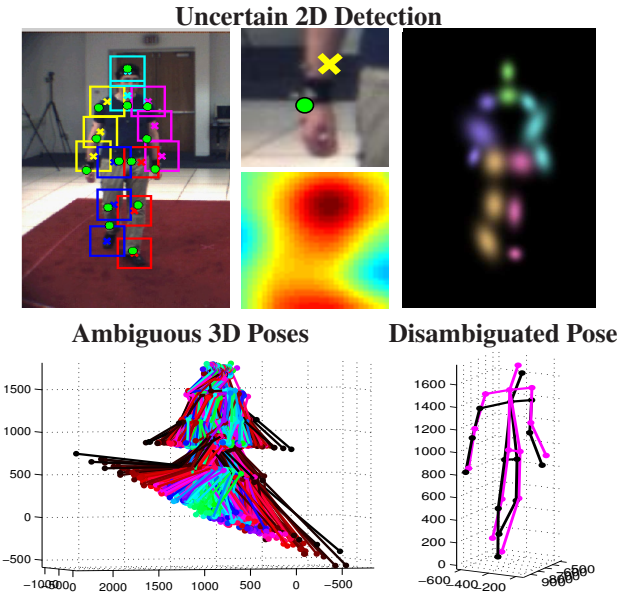


Figure 1: 3D human pose estimation from noisy observations. **Top:** The left image shows the bounding box results of a body part detector and green dots indicate the true position of the joints. Note, in the middle, how the bounding box centers do not match the joint positions. Using the heat map scores of the classifier we represent the output of the 2D detector by Gaussian distributions, as shown on the right. **Bottom:** Using the distribution of all the joints we sample the solution space and propose an initial set of ambiguous poses. By simultaneously imposing geometric and kinematic constraints that ensure the anthropomorphism, we are able to pick an accurate 3D pose (shown in magenta on the right) very similar to the ground truth (black).

In order to robustly retrieve 3D human poses we propose a new approach in which noisy observations are modeled as Gaussian distributions in the image plane and propagated forward to the shape space. This yields tight bounds on the solution space, which we explore using a probabilistic sampling strategy that guarantees the satisfaction of both

This work has been partially funded by Spanish Ministry of Economy and Competitiveness under projects PAU+ DPI2011-27510 and Consolider Ingenio CSD2007-00018; and by the EU project IntellAct FP7-269959. A. Ramisa is supported by a CSIC JAE-DOC grant co-financed by FSE.

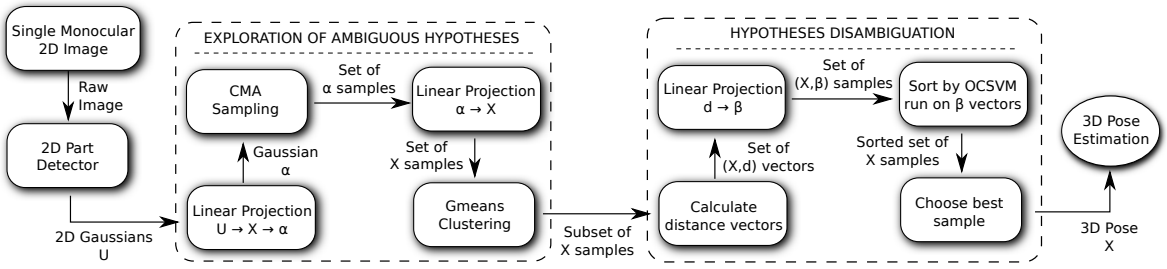


Figure 2: Flowchart of the method presented in this paper for obtaining 3D human pose from single images.

geometric and anthropomorphic constraints. To favor efficiency, the exploration is performed hierarchically, starting from relatively lax and computationally efficient constraints up to highly restrictive and costly ones, until one single shape sample is retained. As we will show in the experimental section, our methodology outperforms approaches that optimize using only geometric constraints.

Overall, we believe our work bridges the gap between current high level 2D detectors [2, 9, 31, 35, 37], and low level geometry-based approaches for 3D pose and shape estimation [10, 18, 19, 20, 26, 33]. The former have shown impressive results in the 2D detection of body parts under harsh conditions, although their resulting accuracy is not optimal. The latter have been proven robust to retrieve pose and shape when accurate observations of the image features are provided. Our approach lies in between both methodologies and estimates the best fitting pose by taking into account not only geometric and kinematic constraints, but also the uncertainty in the observations in a unified process.

2. Related Work

Monocular 3D human pose estimation has generated a wide body of literature [16, 23]. It is a highly ambiguous problem that requires introducing additional knowledge to restrict the size of the solution space. A common approach is to represent the set of pose configurations by a linear combination of deformation modes learned from training data [4, 5, 27]. More sophisticated dimensionality reduction methods have been proposed to represent non-linearities, such as Isomap [34], Gaussian Mixtures [14], spectral embedding [32] or Gaussian Processes [6, 13, 15, 36, 38]. However, most of these approaches require precise initializations and are meant to work in conjunction with temporal priors in a tracking framework.

In order to retrieve 3D human pose from one single image, most approaches rely on discriminative techniques that learn mappings from image features, such as silhouettes, to 3D poses [1, 8, 21, 22, 25, 30]. Support vector machines, nearest neighbors, regression, mixture of experts, or random forests are some of the techniques used for this purpose. While allowing efficient solutions, these methods typically require large training sets to represent the variability of appearance of different people and viewpoints.

Drawing inspiration from [3, 29] we propose retrieving 3D poses from the 2D body part positions estimated by state-of-the-art detectors [2, 9, 31, 35, 37]. Although these detectors require a much reduced number of training samples, as they individually train each of the parts, they have shown impressive results in a wide range of challenging scenarios. However, their solutions have an associated uncertainty which, combined with the inherent ambiguity of the single view 3D detection, may lead to large errors in the estimated 3D shape. This is addressed in [29] by restricting the method to highly controlled settings, and in [3] by imposing temporal consistency. Other approaches [10, 26, 33] guarantee the single frame solution, but simplify the 2D detection process by either manually clicking the position of the 2D joints or directly using the ground truth values obtained from motion capture systems.

In contrast, the approach we propose naturally deals with the uncertain observations of off-the-shelf body part detectors by modeling the position of each body part using a Gaussian distribution that is propagated to the shape space. This sets bounds on the solution search space, which we exhaustively explore to seek for the 3D pose configuration that best satisfies geometric (reprojection and length) and kinematic (anthropomorphic) constraints. To the best of our knowledge, [7] is the only approach that has previously considered noisy observations, but only those related to the root node and not to all the nodes, as we do. In addition, the mentioned work imposes temporal constraints, while we are able to estimate the 3D pose using one single frame.

3. Methodology

Figure 2 outlines our approach, which can be split into three major parts: 2D part detection, stochastic exploration of ambiguous hypotheses and disambiguation. The 2D body part estimation is based on the state-of-the-art detector [37] which is adapted to our usage by obtaining information from the classifier heatmaps to provide local 2D Gaussian inputs. Following [19], this uncertainty is propagated from the image plane to the shape space, thus reducing the size of the search space. We then use stochastic sampling to efficiently explore this region and propose a set of hypotheses that satisfy both reprojection and length constraints. This set of hypotheses must then be disambiguated

by using some additional criteria. We show that only minimizing the reprojection and length errors does not generally give the best results and propose a new method based on coordinate-free geometry to help disambiguate while ensuring anthropomorphic-like shapes.

3.1. 2D Body Part Detection

For body part detection we used [37] which learns a mixture-of-parts tree model encoding both co-occurrence and spatial relations. Each part is modeled as a mixture of HOG-based filters that account for the different appearances the part can take due to, for example, viewpoint change or deformation. Since the parts model is a tree, inference can be efficiently done using dynamic programming, even for a significant number of parts. The output of the detector is a bounding box for each body part, which we convert to a Gaussian distribution by computing the covariance matrix of the classification scores within the box. This is done because the method we propose below to estimate the 3D pose takes as input probability distributions.

3.2. Estimating Ambiguous Solutions

The Gaussian distributions of each body part will be propagated to the shape space and used to propose a set of 3D hypotheses that both reproject correctly onto the image and retain the inter-joint distances of training shapes. However, due to the errors in the estimation of the 2D part location, there is no guarantee that minimizing these errors will yield the best pose estimate. We will show that this requires applying additional anthropomorphic constraints.

The approach we use to propagate the error and propose ambiguous solutions is inspired in [19], originally applied to non-rigid surface recovery. However, note that dealing with 3D human poses has an additional degree of complexity, because most joints can only be linked to two other joints. In contrast, when dealing with triangulated surfaces, each node is typically linked to six nodes. Therefore, the set of feasible human body configurations is much larger than the set of surface configurations. This will require using more sophisticated machinery such as integrating kinematic constraints within the process.

3.2.1 Linear Formulation of the Problem

We represent the 3D pose by a vector $\mathbf{x} = [\mathbf{p}_1^{\mathrm{T}}, \cdots, \mathbf{p}_{n_v}^{\mathrm{T}}]$, where $\mathbf{p}_i$ are the 3D positions of the skeleton joints. The body part detector estimates the 2D position $\mathbf{u}_i$ of each joint $\mathbf{p}_i$ with an associated covariance matrix $\mathbf{\Sigma}_{\mathbf{u}_i}$. Our goal is to retrieve the 3D joint positions from these observations. This can be seen as the solution of a linear system. Assuming the matrix of internal parameters $\mathbf{A}$ to be known, the projection of $\mathbf{p}_i$ onto $\mathbf{u}_i$ may be written as $w_i[\mathbf{u}_i^{\mathrm{T}} \;\; 1]^{\mathrm{T}} = \mathbf{A}\mathbf{p}_i$, where w_i is a projective scalar. This provides 2 linear constraints on $\mathbf{x}$. We can then express the projection of all joints by

$$\mathbf{M}\mathbf{x} = 0 \;, \tag{1}$$

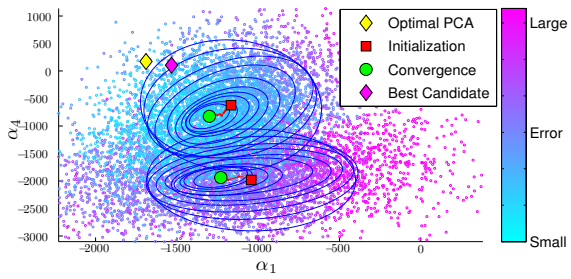


Figure 3: Exploration of the solution space. The figure plots the distribution of samples on the modal weights space and how the covariance matrix is progressively updated using the CMA algorithm. The two distributions represent both hypotheses of the directions the pose can be facing. In addition, the graph depicts the initial and the final configurations obtained with the CMA, and an optimal solution computed by directly projecting the ground-truth pose onto the PCA modes. The *Best Candidate* corresponds to the solution estimated by our approach. Note that although the CMA does not converge close to the optimal solution, some of the samples accumulated through the process lie very close, and thus, are potentially good solutions.

where $\mathbf{M}$ is a $2n_v \times 3n_v$ matrix obtained from the known values $\mathbf{u}_i$ and $\mathbf{A}$. Although minimizing this system may yield to correctly reprojected solutions, there is no guarantee that they resemble a real 3D human pose. This is because $\mathbf{M}$ is rank deficient. We need therefore to include additional constraints. As in most state-of-the-art approaches [4, 5, 27], we will assume we can model the 3D pose as a linear combination of a mean 3D pose $\mathbf{x}_0$ and n_m deformation modes $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_{n_m}]$

$$\mathbf{x} = \mathbf{x}_0 + \sum_{i=1}^{n_m} \alpha_i \mathbf{q}_i = \mathbf{x}_0 + \mathbf{Q}\boldsymbol{\alpha} \;, \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{n_m}]^{\mathrm{T}}$ are the unknown weights that define the current 3D pose. These modes can be obtained by applying Principal Component Analysis (PCA) over a set of pose configurations obtained from the training data. Combining Eqs. (1) and (2), we finally obtain

$$\mathbf{M}\mathbf{Q}\boldsymbol{\alpha} + \mathbf{M}\mathbf{x}_0 = 0 \;. \tag{3}$$

3.2.2 Propagating the Uncertainty to the Shape Space

We must now propagate the 2D Gaussian distributions found on the camera plane to the $\boldsymbol{\alpha}$-subspace of modal weights. Following [19], the mean of this subspace can be computed as the least-squares solution of Eq. (3),

$$\boldsymbol{\mu}_\alpha = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{b} \;, \tag{4}$$

where $\mathbf{B} = \mathbf{M}\mathbf{Q}$ is a $2n_v \times n_m$ matrix and $\mathbf{b} = -\mathbf{M}\mathbf{x}_0$ is a $2n_v$ vector. The components of $\mathbf{B}$ and $\mathbf{b}$ are linear functions of the known parameters $\mathbf{u}_i$, $\mathbf{Q}$ and $\mathbf{A}$. The same can be

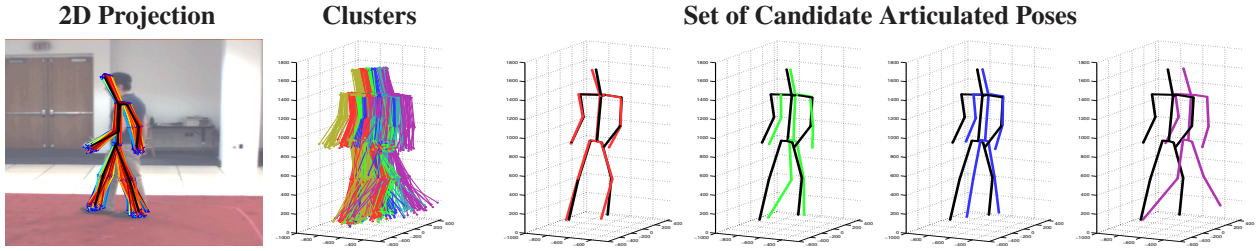


Figure 4: Exploring the space of articulated shapes. The first two plots represent the 2D projection and 3D view of the shape samples we generate. The color of the 3D samples indicates the cluster to which they belong. The four graphs on the right represent the medoids of the clusters, which are taken to be the final set of ambiguous candidate shapes.

done for the $2n_v \times 2n_v$ covariance matrix $\mathbf{\Sigma}_u$ built using the covariances $\mathbf{\Sigma}_{\mathbf{u}_i}$ of each body part. Its propagation yields a $n_m \times n_m$ matrix $\mathbf{\Sigma}_\alpha$ on the modal weights space,

$$\mathbf{\Sigma}_\alpha = \mathbf{J}_B \mathbf{\Sigma}_u \mathbf{J}_B^{\mathrm{T}} \,, \tag{5}$$

where $\mathbf{J}_B$ is the $n_m \times 2n_v$ Jacobian of $(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{b}$.

3.2.3 Proposing Ambiguous 3D Poses

The Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_\alpha, \mathbf{\Sigma}_\alpha)$ represents a region of the shape space containing 3D poses that will most likely project close to the detected 2D joint positions $\mathbf{u}_i$. We will now sample this region and propose a representative set of hypotheses. Note however, that the mean $\boldsymbol{\mu}_\alpha$ computed in Eq. (4) is unreliable, as it is computed from the $\mathbf{u}_i$'s which are not necessarily the true means of the distributions. We therefore do not draw all samples at once. Instead, we propose an evolution strategy in which we draw successive batches by sampling from a multivariate Gaussian whose mean and covariance are iteratively updated using the Covariance Matrix Adaptation (CMA) algorithm [12] so as to simultaneously minimize reprojection and length errors.

More specifically, at iteration k we draw n_s random samples $\{\tilde{\boldsymbol{\alpha}}_i^k\}_{i=1}^{n_s}$ from the distribution $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}^k, \mathcal{M}^2\mathbf{\Sigma}_{\boldsymbol{\alpha}}^k)$, where $\mathcal{M}$ is a constant that guarantees a certain confidence level (we set $\mathcal{M} = 4$ in all experiments). Each sample $\tilde{\boldsymbol{\alpha}}_i^k$ is assigned a weight π_i^k proportional to $\varepsilon_{lr} = \varepsilon_{\mathrm{l}} \cdot \varepsilon_{\mathrm{r}}$, the product of the length and reprojection errors:

$$\varepsilon_{\mathrm{l}} = \sum_{i,j \in \mathcal{N}} \left\| \tilde{l}_{ij} - l_{ij}^{\mathrm{train}} \right\| \sigma_{ij}^{-1}, \tag{6}$$

$$\varepsilon_{\mathrm{r}} = \sum_{i}^{n_v} \sqrt{(\tilde{\mathbf{u}}_i - \mathbf{u}_i)^{\mathrm{T}} \mathbf{\Sigma}_{\mathbf{u}_i}^{-1} (\tilde{\mathbf{u}}_i - \mathbf{u}_i)}, \tag{7}$$

where l_{ij}^{train} is the mean distance in all training samples between the i-th and j-th joints, σ_{ij} is the standard deviation, $\tilde{l}_{ij}$ is the length between joints i and j in the sample $\tilde{\boldsymbol{\alpha}}_i^k$, and the $\tilde{\mathbf{u}}_i$'s are their corresponding 2D projections.

Given the weights π_i^k for all samples, we then update the mean and covariance of the distribution following the CMA strategy. The mean vector $\boldsymbol{\mu}_{\boldsymbol{\alpha}}^{k+1}$ is estimated as a weighted average of the samples. The update of the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\alpha}}^{k+1}$ consists of three terms: a scaled covariance matrix from the preceding step, a covariance matrix that estimates the variances of the best sampling points in the current generation, and a covariance that exploits information of the correlation between the current and previous generations. For further details, we refer the reader to [12].

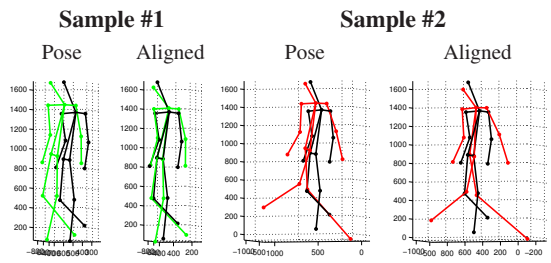
After each iteration a subset of the samples with smaller weights is retained and progressively accumulated for additional analysis. Note that instead of trying to optimize the error function, we use the error function with the CMA optimizer as a way to explore the solution space. When a specific number of samples (10^4 in practice) has been obtained, the problem then becomes how to disambiguate them to find one that represents an anthropomorphic pose.

Orientation Ambiguity. As the detector input does not provide information on the orientation of the subject, we consider the possibility of the pose facing both directions by swapping the detected parts representing the left and right side of the body. This leads to two different distributions which we can then sample from. Figure 3 shows an example of how the solution space is explored. Note that although the CMA algorithm converges relatively far from the optimal solution with minimal reconstruction error, some of the samples accumulated through the exploration process are good approximations. This is the key difference between using a plain CMA, which just seeks for one single solution, and our approach, that accumulates all samples and subsequently uses more stringent –although computationally more expensive– constraints to disambiguate.

Hypotheses Clustering. After exploring the solution space, we have obtained a large number of samples that represent possible poses that have both low reprojection and length errors. However, since many of these samples are very similar, we reduce their number using a Gaussian-means clustering algorithm [11]. As shown in Fig. 4, we then consider the medoid of each cluster to be the candidate ambiguous shape. With this procedure, we can effectively reduce the number of samples from 10^4 to around 10^2.

3.3. Hypotheses Disambiguation

The set of ambiguous shapes has been obtained by imposing relatively simple but computationally efficient con-



	Sample #1	Sample #2
Error Value (ε_{lr})	6.883	6.885
SVM Output	2.8e-04	-7.6e-03
Reconst. Err. (mm)	199.9	214.9
Aligned Err. (mm)	56.4	167.7

Figure 5: The anthropomorphism factor obtained from the OCSVM can be used to choose more human-like models. In the top figures, the black lines represent the ground truth while the colored lines represent the different poses. Note that although Shape #1 is far more human-like than Shape #2, both the error given by $\varepsilon_{lr} = \varepsilon_{\text{l}} \cdot \varepsilon_{\text{r}}$ and the reconstruction error are almost the same. In contrast, the output of the SVM (+1: anthropomorphic; -1:non-anthropomorphic) indicates that Shape #1 resembles more a human-like pose. A good way to validate anthropomorphism is by aligning the pose to the ground truth and measuring the reconstruction error after alignment.

straints based on reprojection and length errors. In this section we will describe more discriminative criteria based on the kinematics of the anthropomorphic pose to further disambiguate them until obtaining a single solution.

For this purpose, we will first propose using a *coordinate-free kinematic representation* of the candidate shapes, based on the Euclidean Distance Matrix. Given the 3D position of the n_v joints, we define the $n_v \times n_v$ matrix $\mathbf{D}$ such that, $d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|$. It can be shown that this representation is unique for a given configuration. In addition, as it is a symmetric matrix with zero entries at the diagonal, it can be compactly represented by the $n_v(n_v-1)/2$ vector

$$\mathbf{d}_{\text{Kin}} = [d_{12}, \cdots, d_{1n_v}, d_{23}, d_{24}, \cdots, d_{(n_v-1)n_v}]^{\text{T}}. \quad (8)$$

Given this unique representation of the pose kinematics, we then propose the treatment of the anthropomorphism as a regression problem. Specifically, we want to be able to calculate how different a 3D pose is from a set of training poses. We deal with this problem by using a one-class Support Vector Machine (OCSVM). The scores computed with this classifier can then be used to distinguish between clusters to determine the most anthropomorphic one.

In order to be able to properly determine the degree of anthropomorphism, and given that we have a limited amount of training data, we need to reduce the size of our pose representation and avoid the curse of dimensionality. For this purpose, we will use again PCA, and we will not directly train the classifier on the whole Euclidean distance vector, but with a linear projection $\boldsymbol{\beta}$ of it.

One important thing to note is that the projection of the distance vectors $\mathbf{d}_{\text{Kin}}$ to the subspace $\boldsymbol{\beta}$ implies a loss of information that can lead to non-anthropomorphic forms being projected close to anthropomorphic forms. In order to account for this effect, it is important to remove the clusters with the worst error value ε_{lr}. As shown in Fig. 5, this increases the likelihood that the results returned by the OCSVM correspond to an anthropomorphic form.

4. Experimental Results

We evaluated the algorithm on two different datasets: the HumanEva dataset [28], which provides ground truth, and the TUD Stadmitte sequence [3], which is a challenging urban environment with multiple people, but without ground truth for a quantitative evaluation.

Regarding the 2D part detector, we used one of the pretrained models included with the software [37], trained on the PARSE dataset [24]. In using these models, we had to deal with an additional source of error, due to the fact that the ground truth joint positions defined in the PARSE dataset and in the HumanEva are not exactly the same. Yet, our approach was robust to this inherent bias.

4.1. Evaluation on the HumanEva dataset

We quantitatively evaluated the performance of our method, using the *walking* and *jogging* actions of the HumanEva dataset. For training the PCA and SVM, we used the motion captured data, independently for each action, for subjects "S1", "S2" and "S3", and used the "validation" sequences for testing.

To speed up evaluation, every 5th frame was used instead of the entire sequence and the average result of 3 repetitions was computed. In order to evaluate our algorithm and not the off-the-shelf 2D part detector, we filtered out the frames where the 2D detector largely failed. This was automatically done by dropping the frames in which there was at least a single part located at more than 80 pixels away from its ground truth.

Fig. 6 shows the distribution of the results on the "S2 walk" sequence. In Fig. 6-Left we plot the average reconstruction error of our approach (*OA*), and compare it against the reconstruction error of *Opt*: the best approximation we could achieve using PCA; *BRec*: sample with minimum reconstruction error among all samples generated during the exploration process; *BRep*: the sample with minimum reprojection error; *BLen*: the sample with minimum length error; and *BErr*: the sample that minimizes $\varepsilon_{lr} = \varepsilon_{\text{l}} \cdot \varepsilon_{\text{r}}$. Note that neither minimizing the reprojection error, the length error nor ε_{lr} guarantees retrieving a good solution. We address this by also maximizing the similarity with anthropomorphic shapes. By doing this, the mean error per joint of

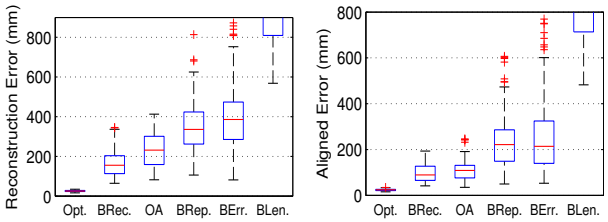


Figure 6: **Left:** Reconstruction errors on the HumanEva dataset for the sequence "S2 walk". **Right:** Same errors after rigid alignment of the shapes with the ground truth poses. It is more representative of the anthropomorphism of the pose compared to the plain reconstruction error, which only considers the distance between the joints of the retrieved pose and the ground truth. See text for a detailed description of the labels.

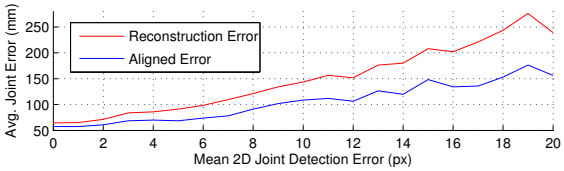


Figure 7: Influence of the 2D detection error on the 3D pose estimation. The size of the Gaussians inputed to our algorithm is maintained constant with $\sigma = 15$ on a single frame of the *walking* sequence of the HumanEva dataset. The mean of the Gaussians defining the 2D body part locations is offset from the ground truth by Gaussian noise of increasing standard deviation. We can see that our algorithm is able to handle large amounts of noise. The values plotted are the average of 100 repetitions.

the shapes we retrieve is around 230mm. Yet, most of this error is due to slight depth offsets which are hard to control due to the noise in the input data. In fact, if we perform a rigid alignment between these shapes and the ground truth ones, the error is reduced to about 100mm (Fig. 6-Right).

Fig. 7 depicts the results of another experiment to show the robustness to noise of our approach. For this purpose, simulated 2D detections with increasing degrees of noise were used to determine how the 2D error propagates to the 3D pose estimation. It can be seen that despite adding large amounts of noise, the 3D pose estimations remain within reasonable bounds.

Finally, numeric results comparing with the state-of-the-art are given in Table 1. Note that this comparison is for guidance only, as different methods train and evaluate differently. For instance, although [6] yields significantly better results, it relies on strong assumptions, such as background subtraction, which both our approach and [3, 7] do not consider. Therefore, we believe that to truly position our approach, we should compare ourselves against [3, 7]. In fact, the performance of all three methods is very similar, but we remind the reader that [3, 7] impose temporal consistency along the sequence, while we estimate the 3D pose using just one single image. A few sample images of the results we obtain are shown in Fig. 8.

4.2. Testing on Street Images

We have also used the TUD Stadtmitte sequence [3] to test the robustness of the algorithm. We consider the scenario with multiple people to detect. This sequence is especially challenging for 3D reconstruction as the camera has a long focal distance, which amplifies the propagation of the 2D errors to the 3D space.

Since we are dealing with real street images, walking pedestrian poses frequently do not match our limited training data: pedestrians may either carry an object or have their hands in their pockets, as seen in Fig. 9. Furthermore, the

	Walking		
	S1	S2	S3
OA	99.6 (42.6)	108.3 (42.3)	127.4 (24.0)
2D Input	14.1 (7.5)	19.1 (8.1)	26.8 (8.0)
[3]	-	107 (15)	-
[7]	89.3	108.7	113.5
[6]	38.2 (21.4)	32.8 (23.1)	40.2 (23.2)
	Jogging		
	S1	S2	S3
OA	109.2 (41.5)	93.1 (41.1)	115.8 (40.6)
2D Input	18.3 (6.3)	18.1 (6.0)	20.9 (6.1)
[6]	42.0 (12.9)	34.7 (16.6)	46.4 (28.9)

Table 1: Comparing the results on the HumanEva dataset for the *walking* and *jogging* actions with all three subjects. All values are in mm with the standard deviation in parentheses if applicable. 2D values are in pixels. Absolute error is displayed for [3, 7], while our approach (OA) and [6] are relative error values. [3, 7] do not provide *jogging* data.

2D body part detector generally fails to find the correct position of the hands (and consequently the arms) because of these occlusions. Despite these difficulties, our method is usually able to find the correct pose.

Analyzing typical failure cases, we can see they all derive from important misdetections. Specifically, a common mistake that our method has trouble recovering from is when the pedestrian is crossing both legs with the feet close together. This occlusion causes the detector to mismatch the feet, and can cause the 3D pose to be estimated facing the opposite direction. More major 2D body part detector failures, such as mixing two nearby pedestrian parts together, can also cause the 3D pose estimation to fail. However, since the output of the OCSVM indicates the anthropomorphism of the estimation, we can use this value to automatically detect failures.

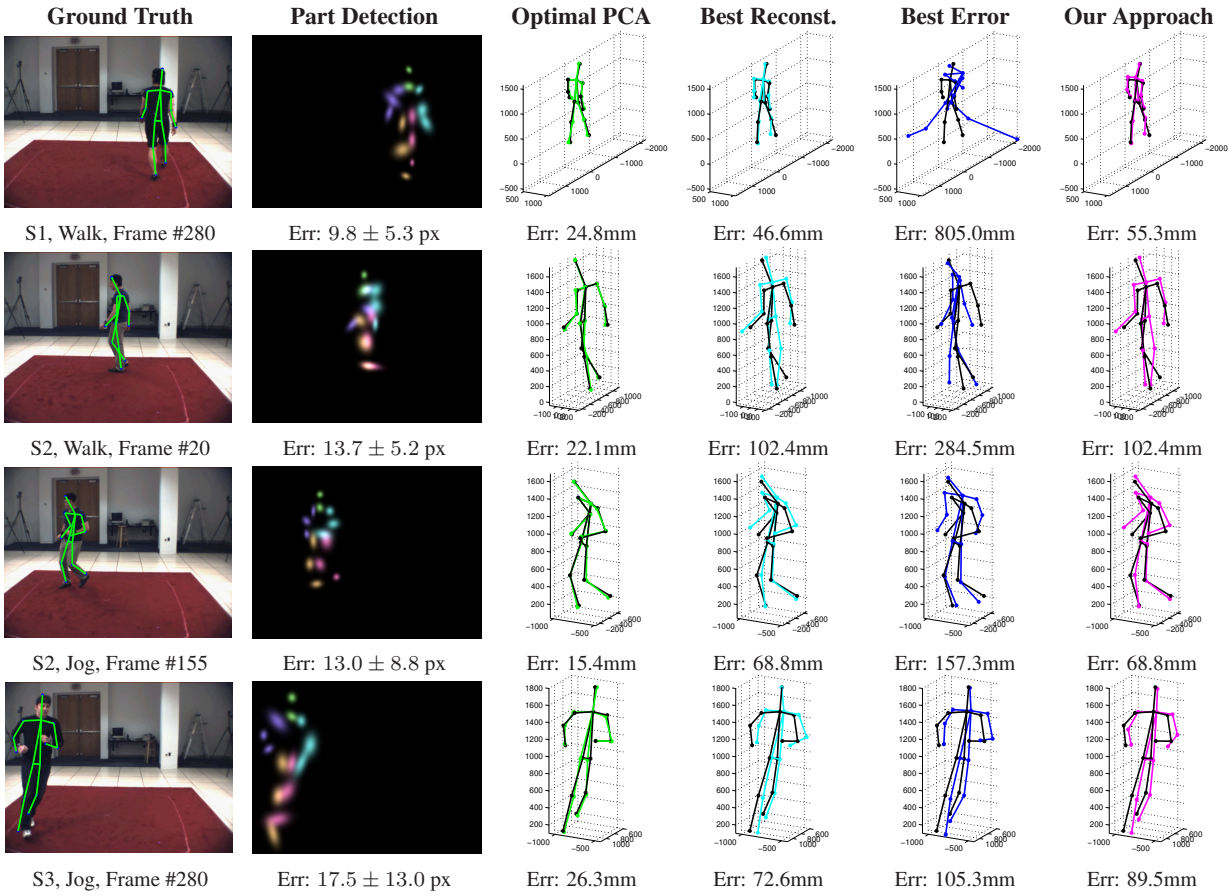


Figure 8: Detection Results. **Leftmost two columns:** Raw image with 2D ground truth projection, and the 2D detection results with the associated average pixel distance from ground truth. **Rightmost four columns:** *Optimal PCA*: projection of the ground truth on the PCA; *Best Reconstruction*: the sample with lowest reconstruction error; *Best Error*: the sample with the lowest error $\varepsilon_{lr} = \varepsilon_{\mathrm{l}} \cdot \varepsilon_{\mathrm{r}}$; and *Our Approach*: the solution obtained. Below each solution we indicate the corresponding reconstruction error (in mm). Note that minimizing ε_{lr} does not guarantee retrieving a good solution.

5. Discussion and Conclusions

In this work we have addressed the ill-posed problem of computing the human 3D pose from a single image, taking as input the noisy predictions of a state-of-the-art 2D body part detector. The uncertainty in the 2D part detection is propagated from the image plane to the 3D shape space, where kinematic constraints are used to disambiguate among the set of feasible 3D shapes. We have found our method to tolerate errors in the 2D localization of the parts of up to 30 pixels. Our results obtained on the HumanEva dataset compare well to those of recent tracking-based approaches that use temporal consistency. Furthermore, we have shown that our method improves significantly if we perform an alignment step, to focus the evaluation on the body pose estimation task rather than on 3D localization. Finally, satisfactory qualitative results have been obtained on an independent, challenging dataset (TUD).

Future work includes training on a large variety of databases, and testing on more independent and "wild" situations. Using the output of our algorithm to feed-back and improve the performance of the 2D detector and exploiting recent non-rigid descriptors [17] is also part of future research.

References

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

[2] M. Andriluka, S. Roth, B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[3] M. Andriluka, S. Roth, B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[4] A. Balan, L. Sigal, M. Black, J. Davis, H. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007.

[5] M. Black, A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *IJCV*, 26(1): 63–84, 1998.

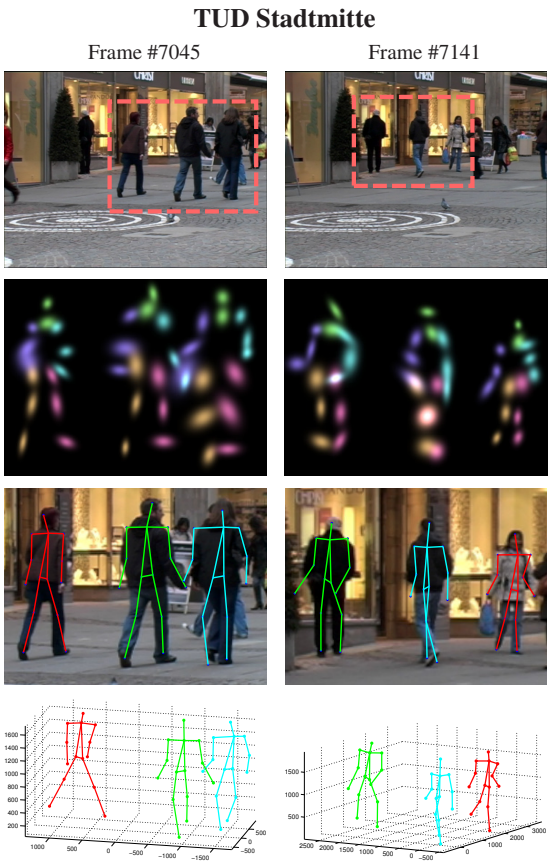


Figure 9: Results on the TUD Stadmitte sequence. **Top three rows:** Raw input image, followed by the detected Gaussians and the reprojection of the estimated 3D pose on the scene. **Bottom row:** Estimated 3D pose.

[6] L. Bo, C. Sminchisescu. Twin Gaussian Processes for Structured Prediction. *IJCV*, 87(1-2): 28–52, 2010.

[7] B. Daubney, X. Xie. Tracking 3d human pose with large root node uncertainty. In *CVPR*, 2011.

[8] A. Elgammal, C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR*, 2004.

[9] P. Felzenszwalb, D. McAllester, D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[10] P. Guan, A. Weiss, A. Balan, M. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009.

[11] G. Hamerly, C. Elkan. Learning the k in k-means. In *NIPS*, 2003.

[12] N. Hansen. The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation. Adv. on estimation of distribution alg.*, pp 75–102. Springer, 2006.

[13] S. Hou, A. Galata, F. Caillette. Real-time body tracking using a gaussian process latent variable model. In *ICCV*, 2007.

[14] N. Howe, M. Leventon, W. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. In *NIPS*, 1999.

[15] N. Lawrence, A. Moore. Hierarchical gaussian process latent variable models. In *ICML*, 2007.

[16] T. B. Moeslund, A. Hilton, V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2): 90–126, 2006.

[17] F. Moreno-Noguer. Deformation and illumination invariant feature point descriptor. In *CVPR*, 2011.

[18] F. Moreno-Noguer, J. Porta. Probabilistic simultaneous pose and non-rigid shape recovery. In *CVPR*, 2011.

[19] F. Moreno-Noguer, J. Porta, P. Fua. Exploring ambiguities for monocular non-rigid shape estimation. In *ECCV*, 2010.

[20] F. Moreno-Noguer, M. Salzmann, V. Lepetit, P. Fua. Capturing 3d stretchable surfaces from single images in closed form. In *CVPR*, 2009.

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

[22] R. Okada, S. Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. In *ECCV*, 2008.

[23] R. Poppe. Vision-based human motion analysis: An overview. *CVIU*, 108(1-2): 4–18, 2007.

[24] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2007.

[25] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, P. Torr. Randomized trees for human pose detection. In *CVPR*, 2008.

[26] M. Salzmann, R. Urtasun. Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In *CVPR*, 2010.

[27] H. Sidenbladh, M. Black, D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, 2000.

[28] L. Sigal, A. Balan, M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, 2006.

[29] L. Sigal, M. Black. Predicting 3d people from 2d pictures. In *AMDO*, 2006.

[30] L. Sigal, R. Memisevic, D. Fleet. Shared kernel information embedding for discriminative inference. In *CVPR*, 2009.

[31] V. Singh, R. Nevatia, C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *ECCV*, 2010.

[32] C. Sminchisescu, A. Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *ICML*, 2004.

[33] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU*, 80(3): 349–363, 2000.

[34] J. Tenenbaum, V. Silva, J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500): 2319–2323, 2000.

[35] T. Tian, S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR*, 2010.

[36] R. Urtasun, D. Fleet, P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, 2006.

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

[38] X. Zhao, Y. Fu, Y. Liu. Human motion tracking by temporal-spatial local gaussian process experts. *IP*, 20(4): 1141–1151, 2011.