A Joint Model for 2D and 3D Pose Estimation from a Single Image

E. Simo-Serra[1], A. Quattoni[2], C. Torras[1], F. Moreno-Noguer[1]
[1]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
[2]Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

Abstract

We introduce a novel approach to automatically recover 3D human pose from a single image. Most previous work follows a pipelined approach: initially, a set of 2D features such as edges, joints or silhouettes are detected in the image, and then these observations are used to infer the 3D pose. Solving these two problems separately may lead to erroneous 3D poses when the feature detector has performed poorly. In this paper, we address this issue by jointly solving both the 2D detection and the 3D inference problems. For this purpose, we propose a Bayesian framework that integrates a generative model based on latent variables and discriminative 2D part detectors based on HOGs, and perform inference using evolutionary algorithms. Real experimentation demonstrates competitive results, and the ability of our methodology to provide accurate 2D and 3D pose estimations even when the 2D detectors are inaccurate.

1. Introduction

Estimating the 3D human pose using a single image is a severely under-constrained problem, because many different body poses may have very similar image projections. In order to disambiguate the problem, one common approach is to assume that an underlying deformation model is available. Linear models [5] or sophisticated dimensionality reduction methods have been used for this purpose [13, 26, 28]. Alternatively, other techniques have focused on learning the mapping from 2D image observations to 3D poses [16, 20, 23]. In any event, most of these generative and discriminative approaches rely on the fact that 2D features, such as edges, silhouettes or joints may be easily obtained from the image.

In this paper, we get rid of the strong assumption that data association may be easily achieved, and propose a novel approach to jointly detect the 2D position and estimate the 3D pose of a person from one single image acquired with a calibrated but potentially moving camera. For this purpose we formulate a Bayesian approach combining a generative latent variable model that constrains the space of possible 3D body poses with a HOG-based discriminative model that constrains the 2D location of the body parts. The two models are simultaneously updated using an evolutionary strategy. In this manner 3D constraints are used to update image evidence while 2D observations are used

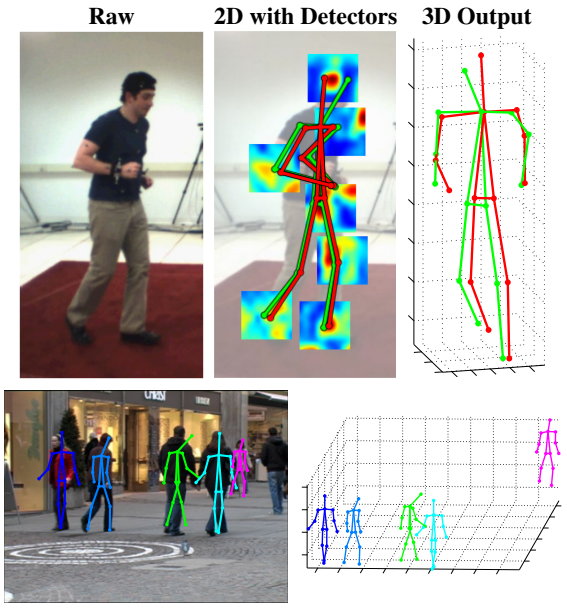

Figure 1: **Simultaneous estimation of 2D and 3D pose.** Top left: Raw input image. Top middle: Ground truth 2D pose (green) and the result of our approach (red). Additionally, we plot a few part detectors and their corresponding score,used to estimate 3D pose. Reddish areas represent regions with highest responses. Top right: 3D view of the resulting pose. Note that despite the detectors not being very precise, our generative model allows estimating a pose very close to the actual solution. Below we show an example of a challenging scene with various pedestrians.

This work has been partially funded by Spanish Ministry of Economy and Competitiveness under projects PAU+ DPI2011-27510, CINNOVA 201150E088 and 2009SGR155; by the EU project IntellAct FP7-ICT2009-6-269959 and by the ERA-Net CHISTERA project VISEN.

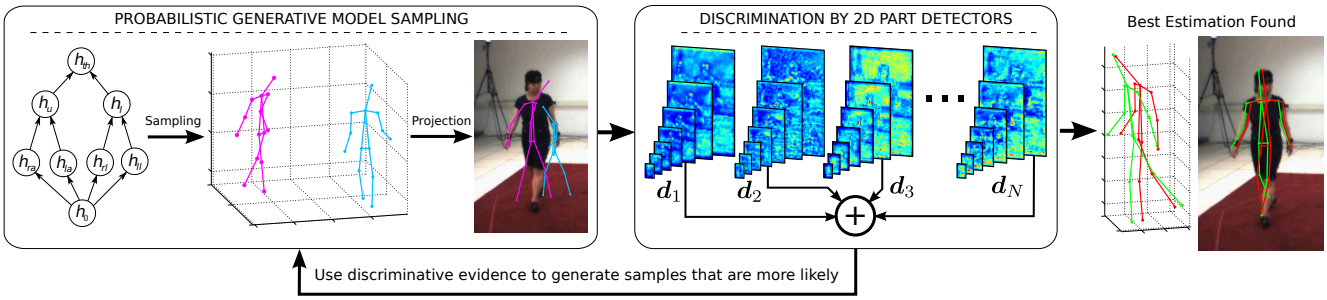

Figure 2: **Method overview.** Our approach consists of a probabilistic generative model and a set of discriminative 2D part detectors. Our optimization framework simultaneously solves for both the 2D and 3D pose using an evolutionary strategy. A set of weighted samples are generated from the probabilistic generative model and are subsequently reweighted by the score given by the 2D part detectors. This process is repeated until convergence of the method. The rightmost figure shows results at convergence where the red shapes are the estimated poses and the green ones correspond to the ground truth.

to update the 3D pose. As shown in Fig. 1 these strong ties make it possible to accurately detect and estimate the 3D pose even when image evidence is very poor.

We evaluate our approach numerically on the HumanEva dataset [22] and qualitatively on the TUD Stadtmitte sequence [3]. Results are competitive with the state-of-the-art despite our relaxation of restrictions as we do not use any 2D prior and instead use directly raw detector outputs.

2. Related Work

Without using prior information, monocular 3D human pose estimation is known to be an ill-posed problem. In order to be disambiguated, many methods to favor the most likely shapes have been proposed.

One of the most straightforward approaches consists of modeling the pose deformations as linear combinations of modes learned from training data [5]. Since linear models are prone to fail in the presence of non-linear deformations, more accurate dimensionality reduction approaches based on spectral embedding [26], Gaussian Mixtures [11] or Gaussian Processes [13, 28, 30] have been proposed. However, these approaches rely on good initializations, and therefore, they are typically used in a tracking context. Other approaches follow a discriminative strategy and use learning algorithms such as support vector machines, mixtures of experts or random forest to directly learn the mappings from image evidence to the 3D pose space [1, 16, 20, 23].

Most of the aforementioned solutions, though, oversimplify the 2D feature extraction problem, and typically rely on background subtraction approaches or on the fact that image evidence, such as edges or silhouettes, may be easily obtained from an image, or even assume known 2D [18, 21].

With regard to the problem of directly predicting 2D poses on images, we find that one of the most successful methods is the pictorial structure model [9] (later extended to the deformable parts model [8]), which represents objects as a collection of parts in a deformable configuration and allows for efficient inference. Modern approaches detect each individual part using strong detectors [2, 25, 27, 29] in order to obtain good 2D pose estimations. The deformable parts model has also been extended to use 3D models for 3D viewpoint estimation of rigid objects [17].

Recently, the estimations of an off-the-shelf 2D detector [29] have already been used for 3D pose estimation in [24]. Yet, and in contrast to the solution we propose here, the 2D estimations are not updated while inferring the 3D shape, and thus, the final 3D pose strongly depends on the result of the 2D detector. The same applies for [3], which computes 2D and 3D pose in two consecutive steps with no feedback. In addition, this work is applied in a 3D tracking domain where temporal constraints play an important role.

3. Joint Model for 2D and 3D Pose Estimation

Figure 2 shows an overview of our model for simultaneous 2D people detection and 3D pose estimation. It consists of two main components, a 3D generative kinematic model, which generates pose hypotheses, and a discriminative part model, which weights the hypotheses based on image appearance. Drawing inspiration from the approach proposed in [2] for 2D articulated shapes, we represent this model using a Bayesian formulation.

With this purpose, we represent 3D poses as a set of N connected parts. Let $L = \{\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N\}$ be their 2D configuration with $\boldsymbol{l}_i = (u_i, v_i, s_i)$. (u_i, v_i) is the image position of the center of the part, and s_i a scale parameter which will be defined below. In addition let $D = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N\}$ be the set of image evidence maps, i.e., the maps for every part detector at different scales and for the whole image. Assuming conditional independence of the evidence maps given L, and that the part map $\boldsymbol{d}_i$ only depends on $\boldsymbol{l}_i$, we have that the likelihood of the image evidence given a specific body

configuration is:

$$p\left(D \mid L\right) = \prod_{i=1}^{N} p\left(\boldsymbol{d}_i \mid \boldsymbol{l}_i\right). \tag{1}$$

In [2], Eq. (1) is further simplified under the assumption that the body configuration may be represented using a tree topology. This yields an additional efficiency gain, as it introduces independence constraints between branches, e.g., the left arm/leg does not depend on the right arm/leg. Yet, this causes the issue of the double counting, where the same arm/leg is considered to be both the left and right one. In [4] this is addressed by first solving an optimal tree and afterwards attempting to correct these artefacts using loopy belief propagation. Instead of using two stages, we directly represent our 3D model using a Directed Acyclic Graph, which enforces anthropomorphic constraints, and helps preventing the double counting problem.

Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be the 3D model that projects on the 2D pose L, where $\boldsymbol{x}_i = (x_i, y_i, z_i)$ is the 3D position of i-th part center. We write the posterior of X given the image evidence D by:

$$p\left(X \mid D\right) \propto \prod_{i=1}^{N} \left(p\left(\boldsymbol{d}_i \mid \boldsymbol{l}_i\right) p\left(\boldsymbol{l}_i \mid \boldsymbol{x}_i\right)\right) p\left(X\right).$$

In order to handle the complexity of directly modeling $p\left(X\right)$, we propose approximating X trough a generative model based on latent variables H. This allows us to finally write the problem as:

$$p\left(X \mid D\right) \propto \underbrace{p\left(H\right) p\left(X \mid H\right)}_{\text{generative}} \underbrace{\prod_{i=1}^{N} \left(p\left(\boldsymbol{d}_i \mid \boldsymbol{l}_i\right) p\left(\boldsymbol{l}_i \mid \boldsymbol{x}_i\right)\right)}_{\text{discriminative}}$$

where the discriminative and generative components become clearly separated. We will next describe each of these components.

3.1. Discriminative Detectors

Recent literature proposes two principal alternatives of discriminative detectors: the shape context descriptors built by applying boosting on the limbs [2], and the HOG template matching approach [29]. For our purposes we have found the HOG-based template matching to be more adequate because it matches our joint-based 3D model better as we can place a detector at each joint part instead of having to infer the limb positions from the joints. In addition, the HOG template matching yields smoother responses, which is preferable when doing inference.

As mentioned above, each part $\boldsymbol{l}_i$ has an associated scale parameter s_i. This parameter is used to pick a specific scale among the evidence maps. Intuitively, if a part is far away, it should be evaluated by a detector at a small scale (high resolution). We therefore approximate the scale s_i as:

$$s_i^{-1} = \alpha^{-1} \beta z_i \tag{2}$$

where α is the focal length of the camera and is used for normalization purposes. The parameter β will be learned off-line and used during inference. Note that despite that this parameter is constant, the scale at which each part is evaluated is different as it depends on its z_i coordinate.

Let $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N\}$ be the set of templates in the HOG space associated to each body part. These templates are provided by [29][1]. Given a body part $\boldsymbol{l}_i$, its image evidence $\boldsymbol{d}_i$ is computed by evaluating the template $\boldsymbol{w}_i$ over the entire image for a range of scales s_i. Fig. 3 illustrates the response of three part detectors at different scales. By interpreting each detector as a log-likelihood, the image evidence of a configuration L can be computed as:

$$\log p\left(L \mid D\right) \approx \text{score}(L) = \sum_{i=1}^{N} k_i \boldsymbol{d}_i(u_i, v_i, s_i)\ , \tag{3}$$

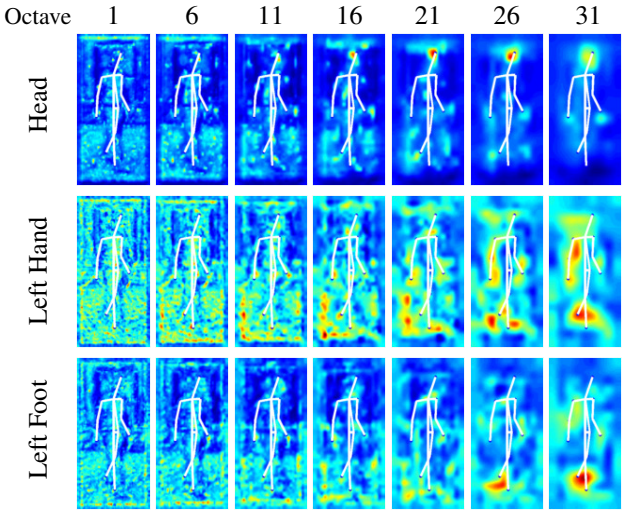

Figure 3: **2D Part Detector**. We visualize the response at different octaves for three different part detectors. We overlay the projection of the 3D ground truth in white to give significance of the accuracy. The outputs are normalized for visualization purposes, with the dark blue and bright red areas corresponding to lower and higher responses respectively. Note that while some detectors, such as the head one in the first row, generally give good results, others do not give good results, such as the left hand detector in the middle row. Our approach can handle these issues by combining these 2D part detectors with a generative model.

[1]Indeed, each of the part detectors provided by [29] are formed by several templates and we use their maximum score for each coordinate (u, v, s). For ease of explanation we refer to them as a single template.

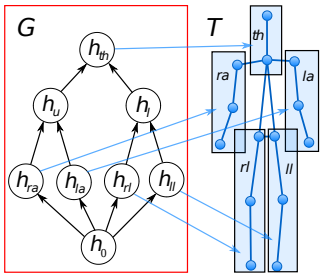

Figure 4: We use a probabilistic graphical model with latent variables G to represent the possible 3D human pose T from a large set of discrete poses. Latent variables can either be mapped to the conjoint motion of various parts or be used as internal states containing internal structure of the pose.

where k_i is a weight associated to each detector, which we learn offline. It is meant to adapt the 2D detectors and 3D generative model due to the fact they were trained independently on different datasets.

Additionally, when evaluating a part detector at a point, we consider a small window from which we use the largest detector value in order to give additional noise tolerance to the detector. We find this necessary as small 2D errors can have large consequences in 3D positioning.

3.2. Latent Generative Model

Our goal is to learn a compression function: $\phi(X^L) : \mathcal{X}^L \rightarrow \mathcal{H}$ that maps points in the high dimensional local 3D pose space $\mathcal{X}^L$ to a lower dimensional space $\mathcal{H}$. The local pose space consists of aligning the coordinates $\mathcal{X}$ to a local reference so that they are independent of the global position in the world and represent only local deformation. We wish to build this compression function so that it efficiently approximates functions of $\mathcal{X}^L$ by sampling $\mathcal{H}$. To create the compression function we define a latent variable model of the joint distribution of a set of variables $H \in \mathcal{H}$ and a pose $X^L \in \mathcal{X}^L$.

As a first step we map the $\mathbb{R}^{N\times 3}$ continuous pose space to a discrete domain by doing vector quantization of groups of 3D joint positions. More specifically, we group the joints into five coarse parts: right arm (ra), left arm (la), right leg (rl), left leg (ll) and torso+head (th). Thus, we can map every pose to a discrete vector $T = [ra, la, rl, ll, th] \in \mathcal{K}^5$ where ra, la, rl, ll, th are cluster indexes belonging to $\mathcal{K} = \{1, 2, .., k\}$ of the corresponding 3D joint positions.

We will now define a joint distribution over latent variables $H \in \mathcal{H} = \{1, .., n\}$ (where n is the number of latent states), and observed variables T. The model is given by the following generative process:

- Sample a latent state i according to $p(h_0)$
- For all the parts associated with arm and leg positions, sample discrete locations: $< ra, la, rl, ll >$ and states $< j, l, m, n >$ according to the conditional distributions: $p(h_{ra} = j, ra \mid i)$, $p(h_{la} = l, la \mid i)$, $p(h_{rl} = m, rl \mid i)$ and $p(h_{ll} = n, ll \mid i)$
- Sample a pair of latent states: $< q, w >$ (associated with the positions of the upper body and lower body joints) according to $p(h_u = q \mid h_{ra} = j, h_{la} = l)$ and $p(h_l = w \mid h_{rl} = m, h_{ll} = n)$
- Sample a discrete location th and a state r from $p\left(h_{(th)} = r, th \mid h_u = q, h_l = w\right)$

Given this generative model we define the probability of a discrete 3D position $T = [ra, la, rl, ll, th]$ as:

$$\begin{aligned} p(T) &= \sum_{\mathcal{H}} p(T, H) \\ &= \sum_{H} p(h_0)\, p(h_{ra}, ra \mid h_0)\, p(h_{la}, la \mid h_0) \\ &\quad p(h_{rl}, rl \mid h_0)\, p(h_{ll}, ll \mid h_0)\, p(h_u \mid h_{ra}, h_{la}) \\ &\quad p(h_{ll}, ll \mid h_0)\, p(h_u \mid h_{ra}, h_{la}) \\ &\quad p(h_l \mid h_{rl}, h_{ll})\, p(h_{th}, th \mid h_u, hl)\,. \end{aligned}$$

Fig. 4 illustrates the graphical model corresponding to this joint distribution, where the graph G specifies the dependencies between the latent states. Since H is unobserved, Expectation Maximization can be used to estimate the model parameters from a set of training poses. Given that G is a Directed Acyclic Graph we can compute all required expectations efficiently with dynamic programming [12]. Once we have learned the parameters of the model we define our compression function to be:

$$\phi(X^L) = \arg\max_{H}\ p\left(X^L, H\right)$$

and our decompression function to be:

$$\phi^{-1}(H) = \arg\max_{X^L}\ p\left(X^L, H\right).$$

Note that the decompression function is not technically speaking the true inverse of $\phi(X^L)$, clearly no such inverse exists since $\phi(X^L)$ is many to one. However, we can regard $\phi^{-1}(H)$ as a "probabilistic inverse" that returns the most probable pre-image of H. Our compression function maps points in $\mathcal{K}^5$ to points in $\mathcal{H}^8$. For example, when $k = 300$ and $n = 9$ we reduce the search space size from 10^{11} to 10^6.

3.3. Parameter Learning

To reduce the number of parameters we need to learn, we take into account the symmetry within the human body, that is, we give the same parameter values to the left and right sides of the body. This allows us to use only 12 parameters instead of 22. Additionally we set a restriction $\sum k_i = 1$ in order to learn relative weighing and reduce the parameters by one. This leaves a total of 12 parameters to be learnt including the detector scale factor β.

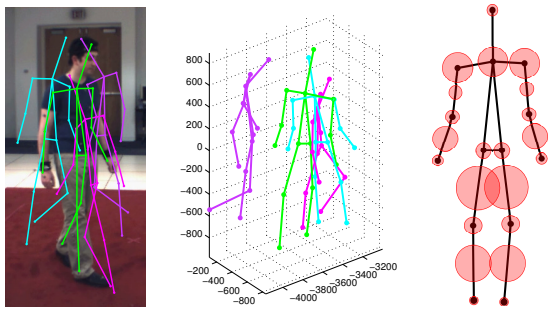

Figure 5: Left and center: Three negative examples used in training. They are coloured by their discriminative score (Eq. (3)) with darker values indicating higher scores. The ground truth is displayed in green. Right: Representation of the trained part weight k_i as a disk with an area proportional to the value. Note that part detectors which one would assume to be most useful, such as the head, have low values. This is caused by the mismatch between the annotations used to train the 2D detectors and the 3D ground truth.

These parameters are learned by translating and rotating random pose samples from the 3D training set and using them as negatives such as those seen in Fig. 5. The parameters are optimized over the difference of the logarithm of expectations of the score from Eq. (3):

$$\arg\max_{\boldsymbol{k},\beta} \log \mathbb{E}\left(\text{score}(L^+)\right) - \log \mathbb{E}\left(\text{score}(L^-)\right)$$

with L^+ and L^- being the sets of positive and negative samples respectively.

3.4. Inference

The inference problem consists of computing:

$$<X^*>= \arg\max_X \prod_{i=1}^{N} p\left(\boldsymbol{d}_i \mid \boldsymbol{l}_i\right) p\left(\boldsymbol{l}_i \mid \boldsymbol{x}_i\right) p\left(X \mid H\right) p\left(H\right).$$

We treat this as a global optimization problem where, given a set of 2D detections corresponding to the different parts from a single image, we optimize over both a rigid transformation and the latent states. Drawing inspiration in [15, 14], we do this using a variation of the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) [10], which is a black box global optimizer that uses a Gaussian distribution in the search space to minimize a function. In our case we perform:

$$\arg\max_{R,\boldsymbol{t},H} \text{score}\left(\underset{R,\boldsymbol{t}}{\text{proj}}\left(\phi^{-1}\left(H\right)\right)\right) + \log\left(p\left(\phi^{-1}\left(H\right), H\right)\right)$$

where $\text{proj}_{R,\boldsymbol{t}}(\cdot)$ is the result of applying a rotation R, a translation $\boldsymbol{t}$ and projecting onto the image plane. We then take $X^*{=}R\phi^{-1}(H^*) + \boldsymbol{t}$, that is, we obtain the most probable X^{L*} given H^* and perform a rigid transformation to the world coordinates to obtain X^*.

States	1	2	3	4	5	**6**	7
UL	123	108	87	64	61	**51**	46
RL	123	107	84	68	77	49	49

Table 1: The influence of the number of latent states per node on the average reconstruction error (in mm). We compare the upper-lower grouping (UL) used in this paper to a right-left grouping (RL), which can be seen to perform roughly the same. The values used are highlighted in bold.

4. Experimental Results

We numerically evaluate our algorithm on the HumanEva benchmark [22], which provides 3D ground truth for various actions. In addition, we provide qualitative results on the TUD Stadmitte sequence [3], a cluttered street sequence. For both cases, we compute the 2D observations using the detectors from [29] trained independently on the PARSE dataset [19]. The 3D model we use consists of 14 joints, each roughly corresponding to one of the 26 detectors from [29]. Additionally, for the limbs we associate an extra detector for a total of 22 detectors (see Fig. 5-right). The latent generative model consists of 8 nodes with 6 latent states. The effect of the number of latent states and structure of the model is shown in Table 1.

4.1. Training

We train our 3D generative model separately for the *walking* and *jogging* actions of the HumanEva dataset using the training sequence for subjects S1, S2 and S3. To avoid overfitting we use part of the training sequence exclusively as a validation set. The part weights k_i and scale factor β are learnt conjointly on the *walking* and *jogging* actions. The same parameters are used regardless of the action as they primarily correspond to the 2D detectors.

4.2. Evaluation

We consider three error metrics: *2D error*, *3D error* and *3D pose error*. The 2D error measures the mean pixel difference between the 2D projection of the estimated 3D shape, and the ground truth. 3D error is the mean euclidean distance, in mm, with the ground truth, and the 3D pose error is the mean euclidean distance with the ground truth after performing a rigid alignment of the two shapes. This error is indicative of the local deformation error. We evaluate three times on every 5 images using all three cameras and all three subjects for both the *walking* and *jogging* actions, for a total of 1318 unique images.

With no additional constraints our optimization framework may take about 30 minutes per frame. This would

Err.	[29] 2D	Ideal Detector 2D	Ideal Detector 3D	Ideal Detector Pose	Our Approach 2D	Our Approach 3D	Our Approach Pose
All	21.7	11.0	106.6	51.6	19.5	237.3	55.3
C1	19.5	11.1	113.8	52.3	18.9	239.1	55.2
C2	22.9	11.1	109.7	51.2	19.6	245.8	55.4
C3	22.8	10.8	96.2	51.2	20.0	227.1	55.4
S1	21.8	10.2	96.8	63.4	19.9	277.2	69.3
S2	21.8	10.8	108.0	44.8	18.6	206.6	46.8
S3	21.6	12.3	119.0	43.7	20.1	221.4	46.6
A1	20.9	10.7	106.0	56.2	19.3	254.4	60.3
A2	22.7	11.3	107.2	46.6	19.7	219.0	50.0

Table 2: Results on the HumanEva dataset for the *walking* (A1) and *jogging* (A2) actions with all subjects (S1,S2,S3) and cameras (C1,C2,C3). We compare with the 2D error obtained using the 2D model from [29], based on the same part detectors we use. 2D, 3D and Pose Errors are defined in the text. Ideal detector corresponds to our approach using Gaussians with 20px covariance as 2D input instead of the 2D detections.

make the task of evaluating all 3×1318 images extremely slow. To speed up the process we have provided a rough initialization to our method. First, instead of considering full images, we crop the original images to have a 60 pixel border around the 2D projection of the 3D ground truth, which on average is 89x288px. Note that this is a criteria which can be easily met with current 2D body detectors.

In addition we have roughly initialized the initial 3D pose parameters with a hyper-Gaussian distribution centered on the ground truth values for R, $\boldsymbol{t}$ and H with the following deviations: $\frac{\pi}{4}$ rad on the rotation around the vertical axis for R; 50 mm deviation in each Cartesian axis for $\boldsymbol{t}$; and 25% of the full latent variable standard deviation for the H. Additionally, we stop the CMA-ES [10] algorithm after 100 iterations. With these assumptions, each 3D pose can be estimated in roughly one minute. Yet, note that while we define the mean to be centered on the ground truth, the CMA-ES algorithm does move in a large area of the search space, as it performs a global and not local optimization. This is shown in Fig. 6. The left column shows an example of optimization process using the constraints just mentioned. Observe that the initial set of explored poses consider very different configurations and space positions.

Table 2 summarizes the results of all experiments. We compare our approach using both Gaussians (20 px Cov.) and the detector outputs as inputs. We see that using ideal detectors, even with large covariances, the absolute error is reduced to 45% of the full approach. An interesting result is that we outperform the 2D pose obtained by [29], using their own part detectors. This can likely be attributed to our joint 2D and 3D model. Nonetheless, although [29] is not an action specific approach as we are, this is still an interesting result as [17] reports performance loss in 2D localization when using a 3D model. Figure 7 shows some specific examples. As expected, performance is better when there are fewer self-occlusions. A full sequence for the *jogging* action with subject S2 is shown in Fig. 8.

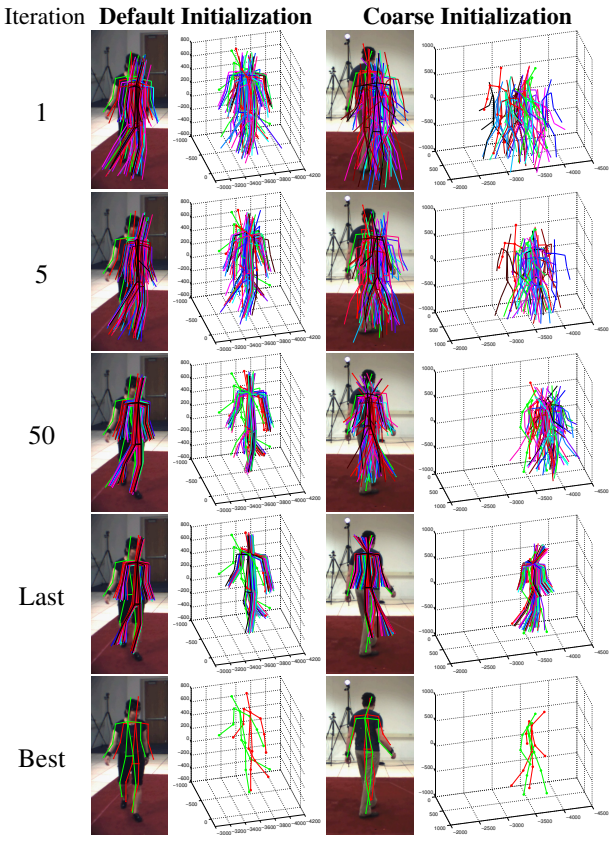

Figure 6: Two different initialization set-ups for our optimization approach. The default initialization consists of optimizing from an initial hyper-Gaussian centered around the ground. The coarse initialization consists in estimating the 3D location based on a 2D pose estimate and using completely random orientation and latent states for the generative model. In the situations in which the detectors are less noisy both initializations perform roughly the same.

We also compare our results with [3, 6, 7, 24]. This is just meant to be an indicative result, as the different methods are trained and evaluated differently. Table 3 summarizes the results using the pose error, corresponding to the "aligned error" in [24]. The two algorithms that use temporal information [3, 7] are evaluated using absolute error. Moreover, [7] uses two cameras, while the rest of the approaches are monocular. Due to our strong kinematic model we outperform all but [6]. Yet, in this work the 2D detection step is relieved through background subtraction processes.

Finally, we present qualitative results on the TUD Stadmitte sequence [3], which represents a challenging real-world scene with the presence of distracting clutter and oc-

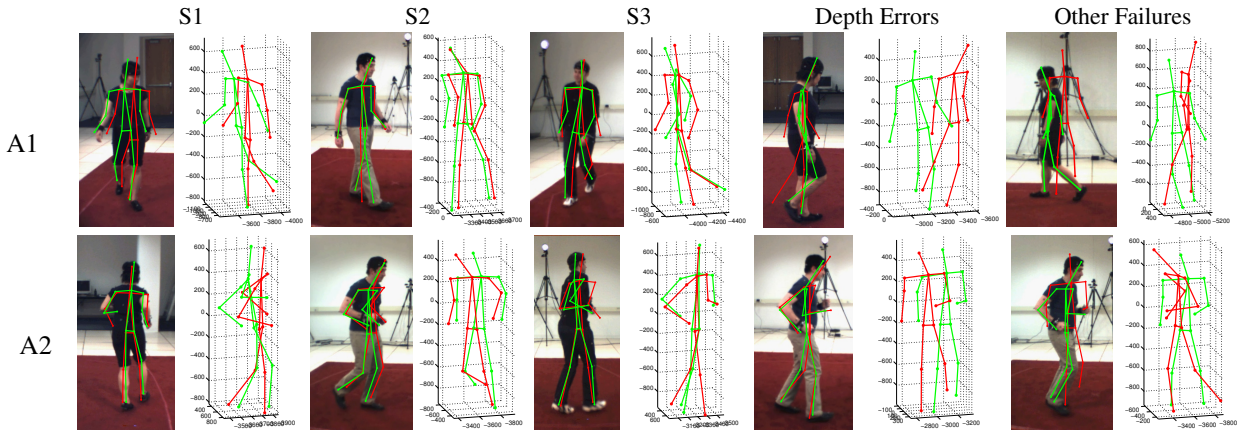

Figure 7: Sample frames from the HumanEva dataset for both the *walking* (A1) and *jogging* (A2) actions. The first three columns correspond to successful 2D and 3D pose estimations for the three subjects (S1,S2,S3). The next two columns show typical failures cases of our algorithm. In the fourth column we see that occasionally we suffer from depth errors, where the 3D pose is correct but its depth is not. In the last column we plot other failures, mostly caused by very large errors of the 2D detector, due to mis-classifications or self-occlusions.

	Walking (A1,C1)		
	S1	S2	S3
Ours	65.1 (17.4)	48.6 (29.0)	73.5 (21.4)
[24]	99.6 (42.6)	108.3 (42.3)	127.4 (24.0)
[7]	89.3	108.7	113.5
[3]	-	107 (15)	-
[6]	38.2 (21.4)	32.8 (23.1)	40.2 (23.2)
	Jogging (A2,C1)		
	S1	S2	S3
Ours	74.2 (22.3)	46.6 (24.7)	32.2 (17.5)
[24]	109.2 (41.5)	93.1 (41.1)	115.8 (40.6)
[6]	42.0 (12.9)	34.7 (16.6)	46.4 (28.9)

Table 3: Comparison against state of the art approaches. We present results for both the *walking* and *jogging* actions for all three subjects and camera C1.

clusions. In addition, since the ground truth is unknown, *no priors on the initialization* of the 3D pose were provided, except the rough bounding boxes for each person (Fig. 6-right). Some results can be seen in Fig. 9. Note that while the global pose seems generally correct, there are still some errors in the 3D pose due to the occlusions and to the fact that the walking style in the wild is largely different from that of the subjects of the HumanEva dataset used to train the generative model.

5. Conclusions and Future Work

The model presented in this paper addresses the ill-posed problem of estimating the 3D pose in single images using a Bayesian framework. We use a combination of a strong kinematic generative model based on latent variables with a set of discriminative 2D part detectors to jointly estimate both the 3D and 2D poses. The results we have obtained are competitive with the state-of-the-art in both 2D and 3D, despite having relaxed the strong assumptions of other methods. Furthermore, the Bayesian framework used is flexible enough to allow extending it further to multi-view sequences, temporal sequences and handle occlusions.

We believe the model we have presented in this paper is a step forward to combining the works of 2D pose estimation with 3D pose estimation. We have shown it is not only possible to estimate 3D poses by applying part detectors used in 2D pose estimation, but that it is also beneficial to the 2D pose estimation itself, as the 2D deformations are being generated by an underlying 3D model.

Future work includes handling occlusions –a weakness of approaches based on 2D detectors–, and improving the handling of rotations by learning a prior distribution and incorporating it in the model. More exhaustive research on different graphical models that can better represent human poses and a deeper analysis of the hyper-parameters chosen are also likely to improve the current method.

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] M. Andriluka, S. Roth, B. Schiele. Discriminative Appearance Models for Pictorial Structures. *IJCV*, 99(3), 2012.

[5] A. Balan, L. Sigal, M. Black, J. Davis, H. Haussecker. Detailed Human Shape and Pose from Images. In *CVPR*, 2007.

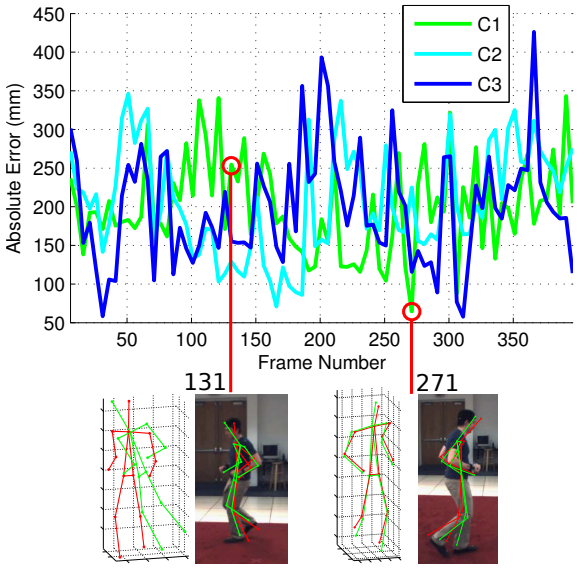

Figure 8: Detection results for subject S2 in the *jogging* action for all three cameras. For this specific action, most of the error comes from only allowing the 3D shape to rotate around the vertical axis. We show two specific frames corresponding to large and small error. The ground truth is displayed in green and the estimated 3D pose in red. On frame 131 we can see that, although the 2D detection is quite accurate, the 3D solution shows large amounts of error due to the strong inclination of the pose. In contrast, frame 271 is accurately detected.

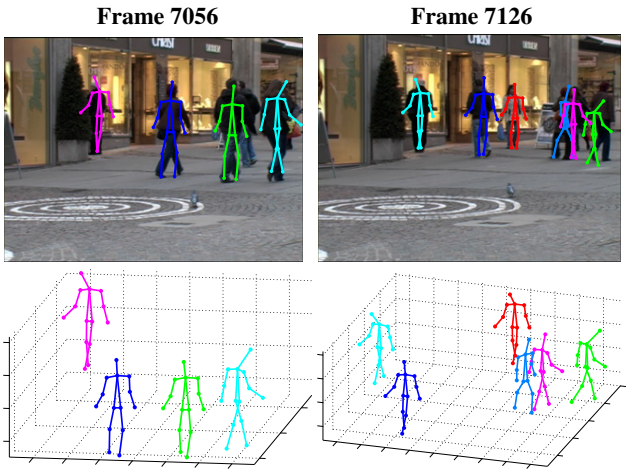

Figure 9: Two sample frames from the TUD Stadtmitte sequence [3].

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

[7] B. Daubney, X. Xie. Tracking 3D Human Pose with Large Root Node Uncertainty. In *CVPR*, 2011.

[8] P. Felzenszwalb, D. McAllester, D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *CVPR*, 2008.

[9] P. Felzenszwalb, D. Huttenlocher. Pictorial Structures for Object Recognition. *IJCV*, 61:55–79, 2005.

[10] N. Hansen. The CMA Evolution Strategy: a Comparing Review. In *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pages 75–102. Springer, 2006.

[11] N. Howe, M. Leventon, W. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. In *NIPS*, 1999.

[12] L. Huang. Advanced Dynamic Programming in Semiring and Hypergraph Frameworks. In *COLING*, 2008.

[13] N. Lawrence, A. Moore. Hierarchical Gaussian Process Latent Variable Models. In *ICML*, 2007.

[14] F. Moreno-Noguer, P. Fua. Stochastic exploration of ambiguities for nonrigid shape recovery. *PAMI*, 35(2):463–475, 2013.

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

[16] R. Okada, S. Soatto. Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images. In *ECCV*, 2008.

[17] B. Pepik, M. Stark, P. Gehler, B. Schiele. Teaching 3D Geometry to Deformable Part Models. In *CVPR*, 2012.

[18] V. Ramakrishna, T. Kanade, Y. Sheikh. Reconstructing 3D Human Pose from 2D Image Landmarks. In *ECCV*, 2012.

[19] D. Ramanan. Learning to Parse Images of Articulated Bodies. In *NIPS*, 2006.

[20] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, P. Torr. Randomized Trees for Human Pose Detection. In *CVPR*, 2008.

[21] M. Salzmann, R.Urtasun. Combining Discriminative and Generative Methods for 3D Deformable Surface and Articulated Pose Reconstruction. In *CVPR*, 2010.

[22] L. Sigal, A. Balan, M. Black. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *IJCV*, 87(1-2):4–27, 2010.

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

[24] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, F. Moreno-Noguer. Single Image 3D Human Pose Estimation from Noisy Observations. In *CVPR*, 2012.

[25] V. Singh, R. Nevatia, C. Huang. Efficient Inference with Multiple Heterogeneous Part Detectors for Human Pose Estimation. In *ECCV*, 2010.

[26] C. Sminchisescu, A. Jepson. Generative Modeling for Continuous Non-Linearly Embedded Visual Inference. In *ICML*, 2004.

[27] T. Tian, S. Sclaroff. Fast Globally Optimal 2D Human Detection with Loopy Graph Models. In *CVPR*, 2010.

[28] R. Urtasun, D. Fleet, P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR*, 2006.

[29] Y. Yang, D. Ramanan. Articulated Pose Estimation with Flexible Mixtures-of-Parts. In *CVPR*, 2011.

[30] X. Zhao, Y. Fu, Y. Liu. Human Motion Tracking by Temporal-Spatial Local Gaussian Process Experts. *Trans. on Image Proc.*, 20(4):1141–1151, 2011.