in: Proc. of REACTS 2011, 1-2 September 2011 Málaga/Spain, pp. 25-39
in conj. with CAIP 2011, 14th Int. Conf. on Computer Analysis of Images and Patterns

Reconstructing The Missing Dimension: From 2D To 3D Human Pose Estimation

Jürgen Brauer and Michael Arens

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation
Gutleuthausstr. 1, 76275 Ettlingen, Germany
{juergen.brauer, michael.arens}@iosb.fraunhofer.de

Abstract. We address the task of estimating a 3D human pose given a 2D pose estimate. A promising approach was presented by Taylor in 2000. The approach needs no learning and is based on a simple principle, namely exploiting the foreshortening information of projected limbs. Though it received only little attention due to two severe restrictions: it uses an unrealistic camera model – the scaled orthographic projection – and yields no unique solution. We show how to overcome both restrictions. We first present an extension of Taylor's original method to a realistic camera model, i.e. perspective projections. Since the method still does not yield an unique solution but a whole set of pose candidates we show how to reduce this set of candidates further by exploiting anatomical constraints and joint angle probabilities. The method is evaluated on the public available TUM kitchen dataset and shows that the average reconstructed joint angle error is in the range of 4.5°-8.2° even for camera views showing strong perspective effects.

Keywords: scene understanding, action recognition, 3D human pose estimation, geometric approach to human pose reconstruction

1 Introduction

Understanding human behavior in image sequences robustly is still an unsolved problem in computer vision. Different approaches have been proposed for recognizing human actions. Some methods learn a mapping from images to action labels directly ([1]), while another approach is first to learn a mapping from images to 3D poses ([2]) and then map 3D poses to action labels ([3]). The problem can be split into even more sub-problems: mapping image evidence descriptors to 2D poses, then mapping 2D poses to 3D poses ([4]), and finally mapping from 3D poses to action labels.

Splitting the problem of action recognition in a divide-and-conquer manner up into several sub-problems has the advantage that different solutions for the sub-problems can be exchanged and tested. For the image → 2D pose estimation

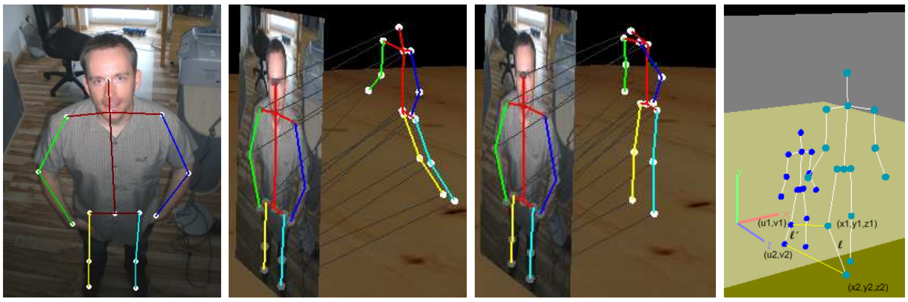

Fig. 1. Approach overview. Left: sample 2D pose with perspective foreshortening effects of limbs. Middle left: 3D pose reconstructed by Taylor's approach that uses a scaled orthographic projection camera model. The strong foreshortening of the upper and lower legs in the camera image can only be explained by a big depth displacement of the corresponding limb start and end points. Middle right: 3D pose reconstructed by our approach that uses a perspective projection camera model and therefore does not need to explain perspective foreshortening effects by wrong depth displacements. Right: Principle used to reconstruct the missing depth based on limb foreshortening information.

sub-problem a huge set of different approaches available is available – often together with reference code (e.g. [5], [6]). The mapping from 2D to 3D poses is currently often learned, e.g. by function regressors such as Gaussian Process Regressors ([7]).

An approach that explicitly models this mapping from 2D to 3D by using knowledge of the 3D to 2D projection process was presented by Taylor [8] in 2000. It is based on a comparison of the actual 3D world limb lengths and the foreshortened projected 2D limb lengths in the image plane in order to reconstruct the missing depth information that does not come with a 2D pose. But Taylor's work experienced only little attention in the pose estimation research community – probably due to two severe restrictions. First, the method assumes that the 2D pose is the result of a scaled orthographic projection of a 3D pose. This is an unrealistic assumption for real world cameras because in such a simple camera model the length of projected limbs do not depend on their distance to the camera. Second, the method results only in a semi-automatic 3D pose reconstruction algorithm since the solution for the depth reconstruction provided is not unique. The method assumes that the user provides the information for each 2D limb which endpoint of the limb is closer to the camera and thereby selects one of many mathematically possible 3D reconstruction solutions.

Mori and Malik [9] approach the problem of the non-uniqueness of the solutions provided by Taylor's original method. A number of example 2D poses is stored in a database. An unknown input image is compared with each stored example 2D pose using shape context descriptor matching. For each 2D pose example in the database the locations of all body parts are stored and additionally the information which limb endpoints are closer to the camera. Thus not only the body part identification information but also this closer limb endpoint information can be transferred automatically to the input image based

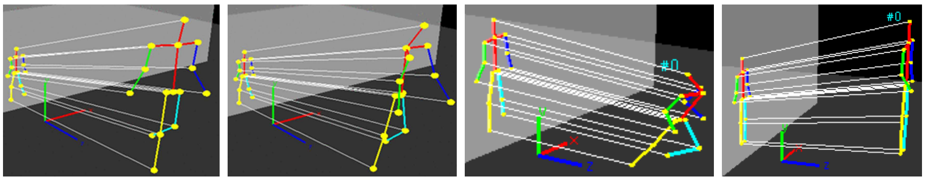

Fig. 2. Limits of Taylor's approach. Left two images: Non-uniqueness. Two different 3D poses reconstructed by Taylor's approach that are projected to the same 2D pose. Right two images: Taking the same scale factor s for a sequence of images will lead to an over- and underestimation of the projection scale s for some frames.

on the shape context descriptor matches. Jiang [10] recently used a simple approach to solve the ambiguity of the 3D pose reconstruction process. The pose candidates are compared to over 4 million poses extracted from the CMU motion capture database. In order to make a comparison of each candidate pose with this huge set of reference poses feasible, poses are split up into upper and lower body poses and are compared separately using an approximative nearest neighbor method. The disadvantage of such an approach is that the method can only recognize poses of actions already stored in the example database. Wei and Chai [11] tackle the problem of estimating the unknown scale parameter in the scaled orthographic projection camera model automatically. Additionally to the bone projection constraints derived by Taylor's original work, the authors establish further constraints based on limb length symmetries and fixed lengths on some rigid subparts of the human body. The 3D pose estimation problem is then formulated as a continuous optimization problem guided by these constraints. Nevertheless, these additional constraints are not sufficient to solve the ambiguity in all cases. Then the pose reconstruction stops and the user has to solve the ambiguity manually before the reconstruction can continue.

Though none of these works tackle the main problem of the unrealistic assumption of a scaled-orthographic camera model. Parameswaran and Chellappa [12] present an 3D pose reconstruction approach that also uses the limb foreshortening information and can deal with perspective projections. First, the possible head orientations are reconstructed by setting up a system of polynomial equations, then the epipolar geometry is recovered, and in a recursive manner the rest of the body joint coordinates are computed using knowledge about the limb lengths. But in their approach the authors have to make two strong assumptions. First, the torso twist has to be small such that the hips and shoulders span up a plane, which means that the approach is not applicable to images of poses for which this is not true. Second, the approach assumes that the locations of four markers on the head are given (e.g. forehead, chin, nose and left or right ear), which is hard to be provided by a 2D pose estimator since it means a fine graded localization.

In this paper we show that we can extend Taylor's approach to perspective projections without having to make such assumptions. The mathematics for the depth reconstruction becomes slightly more complex for the perspective projection case but still tractable. We further cover the question of how to reduce the set of pose candidates without using a too strong bias for certain poses or actions.

In the next section 2 we explain the method. Section 2.1 recapitulates Taylor's original method and shows the limits of this original method. The core contribution of this paper is presented in section 2.2 where we show how to extend Taylor's approach for scaled orthographic projections to perspective projections. Section 2.3 presents a solution for reducing the pose candidate list in order to extract an unique 3D pose estimate. In section 2.4 we explain how to provide estimates for the parameters needed by our reconstruction algorithm for perspective projections and the scale parameter for Taylor's original algorithm. In section 3 we evaluate our 3D pose reconstruction algorithm on the public available TUM kitchen dataset and draw a conclusion based on the results of this evaluation in section 4.

2 Method

2.1 3D Pose Reconstruction for Scaled Orthographic Projections

In 2000 Taylor published an interesting idea in his paper [8]: starting with a 2D pose estimate, we can combine knowledge about the 3D to 2D projection process together with the information how long the projected limbs appear in the 2D image (foreshortening information) in order to reconstruct the missing depth information.

The method assumes that the 2D pose is the result of a scaled orthographic projection of the 3D pose. Thus a 3D point or marker $\boldsymbol{m} = (x, y, z)$ is mapped to a 2D point (u, v) by scaling the x- and y-coordinates by factor s:

$$u = s \cdot x, \quad v = s \cdot y \tag{1}$$

Note that the z-coordinate of the 3D marker has no influence on the resulting projection coordinates in the case of a scaled orthographic projection. This means that the distance of the object to the camera has no importance concerning the projected image. This is of course an oversimplification and wrong for real-world cameras where an object that is far away results in a smaller projected image than an object that is near to the camera. Taylor argues that for cases in which the relative depth of the object of interest is small with respect to the distance between the object and the camera the scaled orthographic projection model is nevertheless appropriate since in such a case to a similar projection like the perspective projection: the small distance differences in z-direction between the limbs are then negligible compared to the distance of the person to the camera.

Assuming we know 1.) the projected coordinates of a limb start $\boldsymbol{m_1} = (u_1, v_1)$ and limb end point $\boldsymbol{m_2} = (u_2, v_2)$, 2.) the limb length l, and 3.) the scale s of the scaled orthographic projection we can reconstruct the displacement $\Delta_z := (z_1 - z_2)$ of the limb in z-direction between two markers $\boldsymbol{m_1} = (x_1, y_1, z_1)$ and $\boldsymbol{m_2} = (x_2, y_2, z_2)$ by reformulating the Euclidian equation as follows. Since $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are projected to points (u_1, v_1) and (u_2, v_2) respectively by a scaled orthographic projection, we have:

$$u_1 - u_2 = s \cdot (x_1 - x_2), \quad v_1 - v_2 = s \cdot (y_1 - y_2) \tag{2}$$

Reformulating the Euclidian equation we get:

$$l^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \tag{3}$$

$$\Leftrightarrow (z1 - z2) = \pm\sqrt{l^2 - (x_1 - x_2)^2 - (y_1 - y_2)^2} \tag{4}$$

$$\Leftrightarrow \mathit{\Delta}_z = \pm\sqrt{l^2 - \frac{(u_1 - u_2)^2 + (v_1 - v_2)^2}{s^2}} \tag{5}$$

Thus we can reconstruct the limb displacement $\mathit{\Delta}_z$ between the markers $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ using the projected points (u_1, v_1), (u_2, v_2), the knowledge about the limb length l and the scale s up to a sign (±) ambiguity. The sign ambiguity stems from the fact that we cannot say whether $\boldsymbol{m_1}$ or $\boldsymbol{m_2}$ is nearer to the camera.

Having two opportunities for reconstructing the relative depth between $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ Taylor's approach gives us $2^{14} = 16384$ possible pose reconstructions for our body model with 14 limbs. Thus the dual ambiguity for a single limb results in an exponential ambiguity for the total 3D pose.

Taylor's approach is interesting since it explicitly models the 2D to 3D reconstruction mapping by exploiting knowledge about the 3D to 2D projection process. It is therefore opposed to approaches that try to model the 2D to 3D reconstruction mapping implicitly, e.g. by training a function regressor with a huge set of (2D,3D) pose pair examples. But it has two severe drawbacks.

First, the method does not provide a unique solution but a set of solutions. In Fig. 2 left we show two examples of 3D poses that stem from the set of 3D pose candidates resulting by Taylor's approach for the same 2D input pose. In Taylor's original work [8] the set of pose candidates was reduced to exactly one by letting the user choose for each limb whether the start or end point of the limb is closer to the camera. In this form it was only a semi-automatic 3D pose reconstruction algorithm. As mentioned in section 1 at least two different ideas were presented in the following years to extend it to a fully automatic 3D pose reconstruction algorithm [9], [10].

Second, the scaled orthographic projection is a poor mathematical description for the projection process of cameras which are better modeled by perspective projections. Thus the reconstructed depths $\mathit{\Delta}_z$ will be wrong in cases where existing perspective effects cannot be modeled by the scaled orthographic projection. Even for cases in which a single image can roughly be approximated by a scaled orthographic projection the approach runs into severe problems since choosing a single scale s for a whole image sequence will result in over- or underestimating the depth $\mathit{\Delta}_z$: if s is overestimated (underestimated) in equation (5), the absolute value of $\mathit{\Delta}_z$ will be overestimated (underestimated) as well. The resulting reconstructed 3D poses are then degenerated and appear exaggerated when the depth is overestimated or too flat when the depth is underestimated. Fig. 2 right shows two examples of 3D poses where the scale was over- and underestimated respectively.

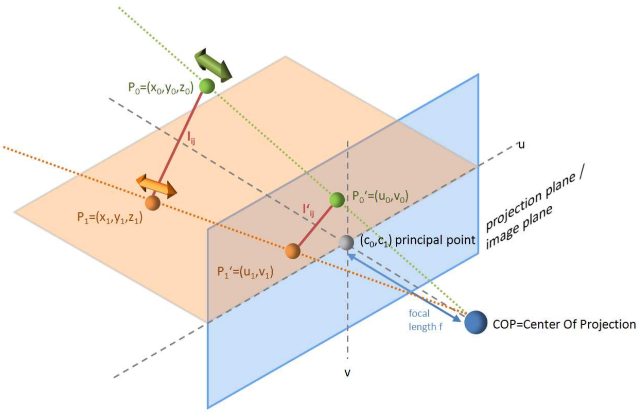

Fig. 3. Perspective projection. 3D markers are projected to 2D points using a perspective projection. Our approach reconstructs a 3D pose step by step. Given an already reconstructed parent marker with coordinates (x_0, y_0, z_0) we reconstruct the child marker coordinates (x_1, y_1, z_1) by using knowledge about the perspective projection and the foreshortened projected limb length l'_{ij} in the image that can be compared with the 3D limb length l_{ij}.

2.2 3D Pose Reconstruction for Perspective Projections

Due to the shortcomings of Taylor's approach resulting from a parallel projection camera model the central question arises whether we can extend Taylor's approach to a more realistic camera model.

A perspective projection projects a 3D point $\boldsymbol{m_i} = (x_i, y_i, z_i)$ to the 2D point $\boldsymbol{m'_i} = (u_i, v_i)$ by

$$u_i = -f\frac{x_i}{z_i} + c_0, \quad v_i = -f\frac{y_i}{z_i} + c_1 \tag{6}$$

where f is the focal length and (c_0, c_1) the origin of the projection plane (principal point) (see Fig.3). Note that now the distance z_i of the point to the image plane has an influence on the resulting projection coordinates.

The principal point (c_0, c_1) is just a 2D translation after scaling the x_i and y_i coordinates by the z_i coordinates and the focal length $-f$ (see equation(6)). When trying to reconstruct the 3D coordinates (x_i, y_i, z_i) based on measured 2D coordinates we can therefore start with inverting the 2D translation step and work with such principal point normalized 2D coordinates $u'_i = u_i - c_0$, $v'_i = v_i - c_1$. We assume that this translation that compensates for the principal point has been done when writing (u_i, v_i) in the following.

The minus sign before the focal length f stems from the fact that the pinhole camera model produces an image of the world that is upside-down. In a code implementation this would mean that we have to rotate the produced 2D image by 180° to view the image. To avoid this we can simply work with f instead of $-f$. If we had an estimate for z_i we could reconstruct x_i, y_i given u_i, v_i, f:

$$x_i = \frac{z_i u_i}{f}, \quad y_i = \frac{z_i v_i}{f} \tag{7}$$

So for each 3D marker $\boldsymbol{m_i}$ to be reconstructed we have one unknown z_i. Since all $\boldsymbol{m_i} = (x_i, y_i, z_i)$ that lie on the perspective projection ray through the point (u_i, v_i) into the direction of the center of projection (COP) are possible candidates for z_i, considering just one point $\boldsymbol{m_i}$ gives us infinite many possible solutions for z_i. Therefore considering the 3D points only isolated does not help. We need to bring in further knowledge by our body model that yields inter-point constraints and thereby constraints the possible values for the $\boldsymbol{m_i}$.

We can think of the points $\boldsymbol{m_i}$ as pearls that can be moved along the perspective projection rays (see Fig. 3). While we move one of these pearls $\boldsymbol{m_i}$ along its projection line, the corresponding projected point $\boldsymbol{m_i'}$ does not change. But since some of these pearls are interconnected and the lengths of these connections are known (relative limb lengths) we can impose further constraints on the relative positions of the pearls. Let us assume that the body structure can be modeled as a kinematic tree with a root marker $\boldsymbol{m_r}$ and l_{ij} denotes the (relative) length of the limb connecting a child marker $\boldsymbol{m_i}$ with its parent marker m_j. The length l_{ij} can be expressed as the Euclidian distance between points $\boldsymbol{m_i}$ and $\boldsymbol{m_j}$:

$$l_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{8}$$

Replacing x_i, x_j, y_i, y_j in equation (8) by equations (7) allows us to transform this Euclidian distance equation into a p-q formula solution (16) of the quadratic equation (14):

$$l_{ij}^2 = (\frac{z_i u_i}{f} - \frac{z_j u_j}{f})^2 + (\frac{z_i v_i}{f} - \frac{z_j v_j}{f})^2 + (z_i - z_j)^2 \tag{9}$$

$$\Leftrightarrow l_{ij}^2 = \frac{1}{f^2}[(z_i u_i - z_j u_j)^2 + (z_i v_i - z_j v_j)^2] + (z_i - z_j)^2 \tag{10}$$

$$\begin{aligned}\Leftrightarrow f^2 l_{ij}^2 &= (z_i u_i)^2 - 2 z_i z_j u_i u_j + (z_j u_j)^2 + (z_i v_i)^2 - 2 z_i z_j v_i v_j + (z_j v_j)^2 + \\ &\quad f^2 (z_i^2 - 2 z_i z_j + z_j^2)\end{aligned} \tag{11}$$

$$\Leftrightarrow f^2 l_{ij}^2 = z_i^2 \underbrace{(u_i^2 + v_i^2 + f^2)}_{=:A} + z_j^2 \underbrace{(u_j^2 + v_j^2 + f^2)}_{=:B} - z_i z_j \underbrace{2(u_i u_j + v_i v_j + f^2)}_{=:C} \tag{12}$$

$$\Leftrightarrow 0 = z_i^2 + \frac{B}{A} z_j^2 - \frac{C}{A} z_i z_j - \frac{f^2 l_{ij}^2}{A} \tag{13}$$

$$\Leftrightarrow 0 = z_i^2 + \underbrace{(-\frac{C z_j}{A})}_{=:p} z_i + \underbrace{(\frac{B}{A} z_j^2 - \frac{f^2 l_{ij}^2}{A})}_{=:q} \tag{14}$$

$$\Leftrightarrow z_{i_{1/2}} = -\frac{p}{2} \pm \sqrt{(\frac{p}{2})^2 - q} \tag{15}$$

$$\Leftrightarrow z_{i_{1/2}} = -\frac{C z_j}{2A} \pm \sqrt{(\frac{C z_j}{2A})^2 - (\frac{B}{A} z_j^2 - \frac{f^2 l_{ij}^2}{A})} \tag{16}$$

We end up with a closed formula (see equation (16)) for the calculation of z_i of a child marker $\boldsymbol{m_i}$ using

- the depth information z_j of a parent marker $\boldsymbol{m_j}$
- the knowledge of the limb length l_{ij}
- the focal length f

– the projected point coordinates $\boldsymbol{m'_i} = (u_i, v_i)$ and $\boldsymbol{m'_j} = (u_j, v_j)$

Intuitively speaking, having already set a parent marker $\boldsymbol{m_j}$ to a fixed position on the perspective projection line, the sign ambiguity before the root in equation (16) corresponds to the fact that we have two possibilities for positioning the child marker $\boldsymbol{m_i}$ on its corresponding perspective projection ray such that the length of the projected line between $\boldsymbol{m_j}$ and $\boldsymbol{m_i}$ fits to the actual measured line between $\boldsymbol{m'_j} = (u_j, v_j)$ and $\boldsymbol{m'_i} = (u_i, v_i)$ (see Fig.3).

Starting with a choice of the z_r coordinate for the root marker $\boldsymbol{m_r}$ we have two possible solutions $z_{i_{1/2}}$ for the first child marker z_i of this root marker according to equation (16). We proceed recursively with this approach within the kinematic tree. Note that in Taylor's original method the depth information for a limb could be reconstructed independently of the depth information for other limbs (compare equation (5). This is no longer true for the the method presented here since in equation (16) the reconstruction of z_i of child marker $\boldsymbol{m_i}$ depends on the reconstructed coordinate of the parent marker $\boldsymbol{m_j}$. By having to consider maximally two possible solutions $z_{i_{1/2}}$ for each marker but the root marker we have to consider maximally 2^L (where L is the number of limbs) possible solutions for the 3D marker positions that all yield the same projection image. As for the scaled orthographic projection scale case the term under the square root in equation (15) has to be positive, i.e. $p^2 - 4q \geq 0$. The determinant $D = p^2 - 4q$ tells us how many solutions are possible for placing the child marker given the parent marker: $D > 0$ will give us 2 possible solutions, $D = 0$ will gives us 1 solution and for $D < 0$ there are no solutions.

2.3 Filtering the Pose Candidates List

For 14 limbs there are maximally 2^{14} mathematically possible 3D pose reconstructions. The ambiguity goes back to the sign ambiguity before the root in equation(16). For reducing the number of pose candidates already during the reconstruction step we check whether any joint angle of the pose violates against anatomical joint constraints. This check is done after each single marker reconstruction step according to equation(16). We exploit anatomical joint constraints given by the right/left knees and right/left elbows: these joints are mainly hinge-joints. Thus only one of the three Euler angles changes within a range of approximately 180°. The interval in which this joint angle for the knee/elbow changes was identified by observing some sample sequences. Observing the corresponding joint angle later outside this interval for a reconstructed pose candidate is considered as a violation of anatomical constraints. Since typically only one of the two z_i solutions is possible for the case of knees and elbows this means that the set of pose candidates also typically reduces to $\frac{2^{14}}{2^4} = 2^{10}$ solutions. Thus the first step to reduce the number of pose candidates is performed already during pose candidate generation.

For the remaining pose candidates we compute a probability for each and choose the one with highest probability. In [10] Jiang uses a database of poses, compares each pose candidate with all poses in the database to find the most similar pose in the database and takes the difference to this most similar pose as a measure for the probability of the pose candidate. But this approach has a severe drawback which is that poses that are not within the database are considered as unlikely or even impossible. The pose candidate that is similar or even equal to the ground truth pose will be assigned only a high probability if there is a similar pose within the database. To avoid such biases to action classes, we first learn the probability $P(\boldsymbol{j_i} = (\alpha, \beta, \gamma))$ of finding a joint in marker $\boldsymbol{m_i}$ in a certain configuration $\boldsymbol{j_i} = (\alpha, \beta, \gamma)$ (the three Euler angles). These probabilities are approximated by relative observation frequencies of 3D example poses (and thereby example joint states) from a motion capture database, as e.g. the CMU motion capture database. While we could define the probability of a reconstructed pose $\boldsymbol{p}$ by the joint probability of all joint angles

$$P(\boldsymbol{p}) = P(\boldsymbol{j_1}, ..., \boldsymbol{j_{N-1}}) \tag{17}$$

this definition again has the drawback that it assigns a probability of 0 to poses not contained in the training data used for estimating P. Therefore we prefer the definition:

$$P(\boldsymbol{p}) = \prod P(\boldsymbol{j_i}) \tag{18}$$

where we assume that the joint states are statistically independent. This definition of pose probability does not impose such a strong prior to poses contained in the training data as definition (17) does. Consider e.g. a motion capture database that contains poses of 1.) persons sitting on a chair and 2.) persons standing and raising their hands, but containing no examples of 3.) persons sitting on a chair and simultaneously raising their hands. Definition (17) will assign a probability of 0 to such latter poses while definition (18) will not. In section 3 we will also analyze whether this definition of probability of a pose is appropriate to find the pose in the set of pose candidates that is most similar to the ground truth pose.

2.4 Estimating the Projection Parameters

Our 3D pose reconstruction algorithm expects as input estimates for the 2D pose, the focal length f, the principal point (c_0, c_1), the limb lengths l_{ij} and the root marker z_r coordinate. The output is the pose candidate with the highest probability.

The 2D input poses are supposed to be provided by a 2D pose estimator (e.g. [13]). By calibrating the camera, the focal length f and the principal point (c_0, c_1) can be computed. Since the principal point is often near to the image center we take it as an estimate for (c_0, c_1). For the limb lengths estimates l_{ij} we first learn the relative lengths r_{ij} of all limbs in terms of the person's size using example 3D poses from motion capture data and then scale the limb lengths to a person of average size, i.e. $l_{ij} = r_{ij} * s$ where s is the person's size estimate and is set to $s = 1692.5$ mm (which is the average size of US adults averaged

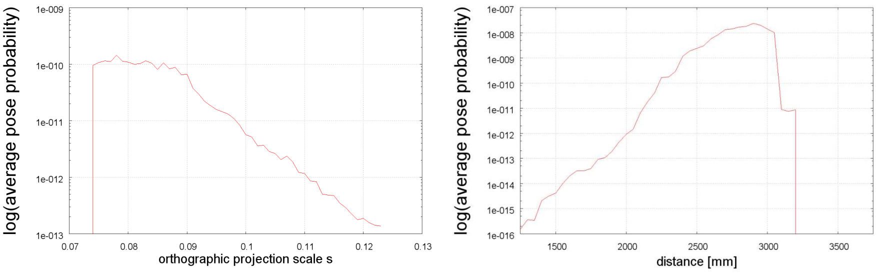

Fig. 4. Left: Average (log) probability of reconstructed poses in dependence of orthographic projection scale estimate s. Right: Average (log) probability of reconstructed poses in dependence of marker distance estimate z_r

over both genders). The computed values r_{ij} were found to fit well to typical limb proportions used for drawing paintings of humans (e.g.: upper and lower leg have the same size).

The z-coordinate z_r for the root marker is estimated by the distance of the person to the image plane. It is possible to use the visual appearance scale of the person to get a rough estimate for this distance since the person will appear smaller if the distance increases but the mapping from this appearance scale to a distance estimate will depend on the camera projection properties. Here we propose another indirect method which showed to yield robustly good distance estimates. The idea is first to reconstruct the set of all possible candidate poses for different sample distances z_r, to compute the average probability of the reconstructed poses at that distances and to choose the one with the highest average pose probability. Fig 4 right shows such a distance / average pose probability plot for a sample frame. The ground truth distance is at approx. 2900 mm where the average pose probability for the set of all reconstructed poses reaches a maximum. This can easily be explained by the fact that for distances different from the ground truth distance, the reconstructed poses often can still be squeezed into the perspective rays bundle but the resulting poses will be degenerated in the sense that the resulting joint angles are unlikely which in return results in poses with low probabilities. We can go even further and choose not only the distance estimate but also the focal length estimate automatically by considering the average probability of the reconstructed poses. This was done e.g. for the image in Fig. 1 where no focal length estimate was available due to missing camera calibration data.

We can use the same idea to extend Taylor's semi-automatic 3D pose estimation approach for scaled orthographic projections to a fully automatic 3D pose estimation approach where the orthographic projection scale estimate s is determined automatically. Since the term under the root in equation(5) can never be negative (since Δ_z has to be a real and not a complex number for real poses) for all limbs in our body model we can first estimate a lower bound for the scale s by $s^\star = \max\{s : s = \sqrt{(u_1-u_2)^2+(v_1-v_2)^2}/l_{ij}\}$ and use this scale estimate as a start value for a search for an estimate for s where we try different scales $s = s^\star + t$ with $t \geq 0$ and reconstruct all possible pose candidates using Taylor's reconstruction equation (5) with such scale estimates s. Then we compute the average probability of the pose candidates in dependence of s and choose the orthographic scale s where we find the pose candidates with the highest average probability. Fig.4 left shows such an example search for a single frame where the best scale ($s = 0.078$) is near to the start search value ($s^\star = 0.074$).

3 Experiments

For evaluating our method we choose the public available TUM kitchen dataset[1] [14] since it provides 3D motion capture ground truth data which allows us to compare our reconstructed poses with the ground truth poses. It consists of 20 action sequences performed by four different persons where in each 1-2 minutes sequence a person lays a table in a kitchen. Video data is provided for four different cameras that are arranged in the four corners of the kitchen.

Exp no.	**test data**	**no of frames**	**subject**	**camera perspective**	**person size estimation error [cm]**
1a	0-0 (cam 0)	439	1	strong	big(10)
1b	0-0 (cam 1)	439	1	strong	big(10)
2a	0-9 (cam 0)	587	4	strong	small(5)
2b	0-9 (cam 1)	587	4	strong	snall(5)
3a	0-0 (cam 2)	439	1	weak	big(10)
3b	0-0 (cam 3)	439	1	weak	big(10)
4a	0-9 (cam 2)	587	4	weak	small(5)
4b	0-9 (cam 3)	587	4	weak	small(5)

Table 1. Experiments definition. We test our 3D pose reconstruction algorithm on sequences with different strong perspective effects and errors in the assumed person size.

Table 1 shows the definition of the experiments conducted. All four cameras are mounted in the top corners of the room. But since camera 0 and 1 (see Fig.5 1a-2b) are near to the cupboard where the subjects take out different objects, the 2D poses recorded from these cameras show strong perspective effects while camera 2 and 3 (see Fig.5 3a-4b) that are at the other end of the room show smaller perspective effects since the persons never come near to these two cameras while the videos are recorded. In experiments 1a/1b/2a/2b we test the reconstruction performance when using 2D poses as input that have these strong perspective effects, while in experiments 3a/3b/4a/4b the perspective effects are less strong. The limb lengths l_{ij} stem from a person with average size (169cm) but subject 1 has a size of 159cm and subject 2 a size of 174cm (computed based on their 3D motion capture data). Thus in experiments 1a/1b/3a/3b the difference between assumed vs. actual size of the person is bigger (10cm) than in experiments 2a/2b/4a/4b (5cm). Since the assumed person size is wrong in all cases, the limb length estimates l_{ij} will be as well, i.e. we can test how good the

[1] `http://ias.cs.tum.edu/download/kitchen-activity-data/`

3D pose reconstruction algorithm can deal with wrong limb length estimates. In total we test our 3D pose reconstruction algorithm on 4104 frames.

Exp no.	**our approach error** [°] (dev)	**extended Taylor error** [°] (dev)	**error cmp**	**our approach error** [**mm**](dev)	**extended Taylor error** [**mm**](dev)	**error cmp**
1a	6.1 (3.2)	6.9 (3.5)	-12%	131.5 (33.1)	152.4 (35.4)	-14%
1b	6.5 (4.0)	7.2 (3.9)	-10%	142.4 (36.9)	160.7 (36.2)	-11%
2a	4.5 (3.7)	6.7 (3.6)	-33%	94.9 (32.9)	134.2 (29.2)	-29%
2b	4.9 (3.6)	7.1 (3.5)	-31%	98.3 (31.6)	139.9 (30.0)	-29%
3a	8.2 (3.5)	7.5 (3.4)	+9%	158.5 (40.9)	150.41 (26.3)	+5%
3b	7.3 (3.7)	7.5 (3.3)	-3%	151.1 (41.5)	157.7 (34.5)	-4%
4a	6.6 (3.7)	6.3 (3.0)	+4%	124.2 (41.2)	129.7 (23.2)	-4%
4b	5.8 (3.6)	6.6 (3.3)	-12%	111.1 (36.6)	131.5 (27.2)	-16%

Table 2. Experimental results. Average error in degree and mm for the final 3D pose estimate for our approach and the extended Taylor approach.

For measuring the quality of a reconstructed 3D pose we compare it with the corresponding 3D ground truth pose. Two error measures are used: the average joint angle difference (specified in degree) and the average marker position difference (specified in mm). In table 2 we present these errors for the final 3D pose estimate (the pose candidate with the highest probability) and also compare it with the extended version of Taylor's approach where we automatically compute an estimate s for the orthographic projection scale for each frame, reconstruct all pose candidates with that scale estimate and choose the pose with the highest probability. In table 3 we additionally specify the errors for the 3D pose candidate in the candidate list that is most similar to the ground truth pose. This gives an impression of which pose error could be reached when exchanging the method for searching a final pose estimate in the list of pose candidates used here by a better method.

Exp no.	**our approach best error** [°](dev)	**extended Taylor best error** [°](dev)	**error cmp**	**our approach best error** [**mm**](dev)	**extended Taylor best error** [**mm**](dev)	**error cmp**
1a	4.3 (1.3)	4.7 (1.2)	-9%	83.8 (29.5)	113.5 (20.9)	-26%
1b	4.0 (1.9)	4.7 (1.7)	-15%	73.4 (32.9)	111.4 (26.9)	-34%
2a	2.7 (1.2)	4.7 (0.9)	-43%	67.5 (27.6)	113.8 (22.0)	-41%
2b	2.8 (1.6)	4.8 (1.3)	-42%	63.0 (22.2)	115.6 (23.8)	-46%
3a	6.0 (1.9)	5.2 (1.4)	+15%	110.4 (45.1)	118.9 (26.4)	-7%
3b	5.1 (1.8)	4.9 (1.4)	+4%	97.4 (35.1)	110.3 (28.9)	-12%
4a	4.7 (2.0)	4.5 (1.2)	+4%	95.2 (42.8)	107.6 (22.8)	-12%
4b	3.9 (1.8)	4.6 (1.4)	-15%	78.1 (27.8)	111.2 (29.4)	-30%

Table 3. Experimental results. Average error in degree and mm for the best pose candidate which is the 3D candidate pose in the set of candidates that is most similar to the ground truth 3D pose.

The results in table 2 show that our 3D pose reconstruction algorithm can reconstruct 3D poses up to an average joint angle error in the order of 4.5°-8.2°

or 9.4cm-15.8cm. Table 3 shows that we could get even better (up to 2.7°-6.0° or 6.3cm-11cm) since we do not manage to find always the best candidate – which is the pose candidate most similar to the ground truth pose – from the set of computed pose candidates. This means that the definition of pose probability in equation (18) could still be improved. Especially if the person is viewed from a strong perspective (Exp. 1a-2b) we can reduce the error in the range of 10%-33% (up to 9%-46%) compared to the extended Taylor approach. In the weak perspective experiments (Exp. 3a-4b) the person is viewed from a perspective close to a scaled orthographic projection. In such situations there is no gain using our method. The error of the final 3D pose estimate will be smaller the better we can estimate the person size (compare Exp.2a/2b/4a/4b to 1a/1b/3a/3b).

For a fair comparison of this quantitative results with work of other authors one has to keep in mind that ground truth 2D input poses were used since the computed 3D pose reconstructing errors were supposed to reflect the quality of our 2D to 3D reconstruction algorithm and not simultaneously the quality of 2D poses provided by a 2D pose estimator. Therefore we compare the resulting angle errors computed here only with work using similar input. To the best of our knowledge the smallest 3D pose estimation errors were presented by Agarwal and Triggs [15]. They used ideal person silhouettes as input which were generated by rendering avatars. The silhouette descriptors were mapped to 3D poses using different regressor approaches as ridge regression, Relevance Vector Machine (RVM) regressors and Support Vector Machine (SVM) regressors with linear and non-linear kernel bases. The smallest reported mean angle error was 5.91° using SVM regression. Averaging over all experiments the average joint angle error using our method is 6.2° (compare table 2 first column). Thus our inverse perspective projection approach yields 3D poses with state-of-the-art quality.

4 Conclusion

We have derived a method for automatic reconstruction of 3D human poses based on 2D input poses and a focal length estimate. In contrast to many other 3D pose estimation approaches our 3D reconstruction algorithm does not only estimate the 3D articulation of a person, but also provides 3D coordinates for all body markers in the camera coordinate system.

Our approach is based on explicitly modeling the 3D to 2D perspective projection and exploits the foreshortening information of body limbs in the camera image in order to recursively reconstruct the missing depth information and thereby lifting 2D poses to 3D pose estimates. In contrast to many approaches that try to learn the 2D to 3D reconstruction mapping without using any knowledge of the perspective projection by using training (2D,3D) pose examples for adjusting some regression function we have presented an approach that allows a clear understanding of how the 3D pose candidates are constructed by doing an inverse perspective projection per limb. The intrinsic ambiguity of the reconstructed poses is solved by using joint angle probabilities while simultaneously keeping care of not using a too strong bias to special actions.

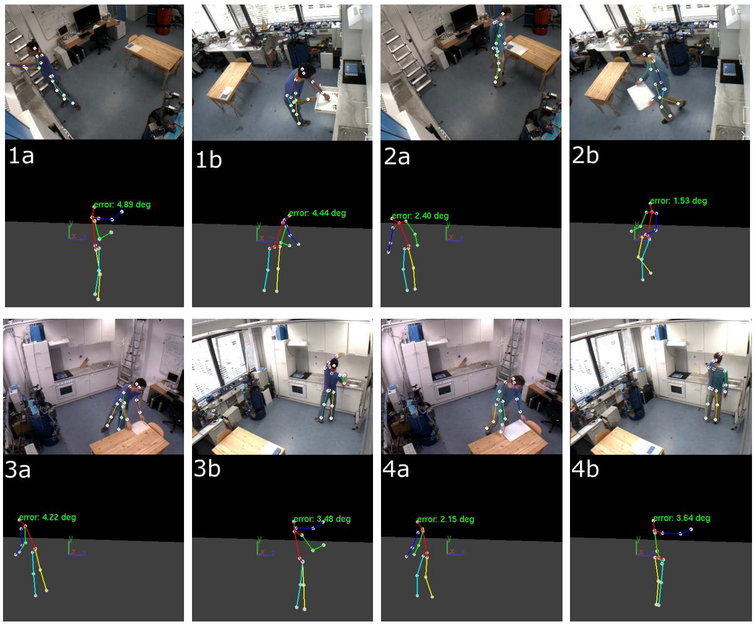

Fig. 5. Examples of reconstructed 3D poses. We present one example of a reconstructed 3D pose for each of the experiments 1a-4b.

In contrast to Taylor's semi-automatic approach the approach presented here is a fully automatic approach capable of reconstructing 3D poses even for 2D input poses that underly strong perspective effects and yields a single final 3D pose estimate. As a side-product a new method was proposed to estimate the person's distance to the image plane based on the average probability of the set of reconstructed poses at different distances. The idea of using the average probability of reconstructed poses for choosing reconstruction parameters was shown to be suitable to estimate the scale parameter in Taylor's approach as well.

Further work will try to quantify the dependence of the quality of reconstructed 3D poses based on the quality of the 2D pose estimates.

References

1. Alexander Kläser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d - gradients. In *Proc. of BMVC*, pages 995–1004, 2008.
2. Mathieu Salzmann and Raquel Urtasun. Implicitly constrained gaussian process regression for monocular non-rigid pose estimation. In *Advances in Neural Information Processing Systems 23*, pages 2065–2073, 2010.
3. Wenjuan Gong, Andrew D. Bagdanov, F. Xavier Roca, and Jordi Gonzàlez. Automatic key pose selection for 3d human action recognition. In *AMDO*, pages 290–299, 2010.
4. Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *Proc. of CVPR 2010*, USA, 2010.

5. V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. of CVPR 2008*, pages 1–8, 2008.
6. Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.
7. R. Urtasun, D.J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *Proc. of ICCV 2005*, pages 403–410, 2005.
8. Camillo J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU*, 80:349–363, 2000.
9. Greg Mori and Jitendra Malik. Recovering 3d human body configurations using shape contexts. *PAMI*, 28(7):1052–1062, 2006.
10. Hao Jiang. 3d human pose reconstruction using millions of exemplars. In *Proc. of 20th ICPR*, pages 1674–1677, 2010.
11. Xiaolin K. Wei and Jinxiang Chai. Modeling 3d human poses from uncalibrated monocular images. In *ICCV*, pages 1873 –1880, October 2009.
12. Vasu Parameswaran and Rama Chellappa. View independent human body pose estimation from a single perspective image. In *Proc. of CVPR 2004*, pages 16–22, Washington, DC, USA, 2004.
13. Jürgen Müller and Michael Arens. Human pose estimation with implicit shape models. In *Proc. of ACM ARTEMIS 2010 - International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream*, Florence/Italy, 2010.
14. Moritz Tenorth, Jan Bandouch, and Michael Beetz. The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS). In conjunction with ICCV conf.*, 2009.
15. A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 28(1):44–58, 2006.