Generative 2D and 3D Human Pose Estimation with Vote Distributions

Jürgen Brauer, Wolfgang Hübner, Michael Arens

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

Abstract. We address the problem of 2D and 3D human pose estimation using monocular camera information only. Generative approaches usually consist of two computationally demanding steps. First, different configurations of a complex 3D body model are projected into the image plane. Second, the projected synthetic person images and images of real persons are compared on a feature basis, like silhouettes or edges. In order to lower the computational costs of generative models, we propose to use vote distributions for anatomical landmarks generated by an Implicit Shape Model for each landmark. These vote distributions represent the image evidence in a more compact form and make the use of a simple 3D stick-figure body model possible since projected 3D marker points of the stick-figure can be compared with vote locations directly with negligible computational costs, which allows to consider near to half a million of different 3D poses per second on standard hardware and further to consider a huge set of 3D pose and configuration hypotheses in each frame. The approach is evaluated on the new Utrecht Multi-Person Motion (UMPM) benchmark with the result of an average joint angle reconstruction error of 8.0°.

1 Introduction

Estimating the 2D and 3D articulation of a person is an important step for action recognition and automatic visual scene understanding. There is a huge amount of literature on human pose estimation. Surveys are provided e.g. by Sminchisescu [1], Poppe [2], and Ji and Liu [3]. Work on human pose estimation can be divided into *top-down* (generative) and *bottom-up* (discriminative, conditional, recognition-based) approaches.

Bottom-up approaches try to predict the 2D or 3D pose directly given image features. There are model-free and model-based bottom-up approaches. Model-free approaches do not make use of a body model, but learn the mapping from

published at 8th Int. Symposium on Visual Computing (ISVC) 2012, 16-18 July, Crete, Greece. This is the author's camera-ready version. The original publication is available at www.springerlink.com

image features to 2D/3D directly using a huge training set of image and 2D/3D pose pairs. In contrast, model-based bottom-up approaches make explicit use of a body model to map the features to a 2D/3D pose.

Top-down approaches are inherently model-based. Hypothesized 3D poses of a human body model are rendered into the 2D image and compared with image features. A key issue for any top-down approach is the definition of a robust matching method, which allows to compare a synthetic view and a real image. For modelling the 3D body different geometric primitives as ellipsoids [4], cylinders [5], or super-quadrics [6] are used. The most detailed 3D human body model so far is [7] and consists of a polymesh based shape model with 25,000 polygons. The authors generate 2D silhouette projections of different 3D pose hypotheses and compare these with the current estimated silhouette from the person image to be analyzed. Beside silhouettes, edges are the image features mostly used as basis for the projected model vs. image evidence comparison. [8] compare edges from the projected model with computed edges by searching into the direction of the projected edge normal for an edge extracted from the image. An alternative approach to using fixed geometric primitives for modeling the 3D body model was recently presented [9]. For each limb a model of its projected shape is learnt using Microsoft's Kinect, which provides a easy way to collect example data, since it provides 3D poses and camera images simultaneously. Generative approaches are claimed to be computationally demanding compared to discriminative approaches due to the high computational costs for projecting a huge set of 3D pose candidates into the image and to compare each projection with the person image using low-level image features as edges and silhouettes.

We propose a computationally efficient approach which avoids the need for a complex 3D body model. An Implicit Shape Model (ISM) ([10], [11]) is used to generated a distribution of votes for potential body landmark locations. Thereby we replace the low-level image evidence with a representation that one the hand, still captures the ambiguity inherent in the 2D domain, but on the other hand, allows to use a simple 3D stick-figure body model, since we can compare projected marker points of this 3D body model with the votes directly and with low computational costs.

There are two approaches ([12], [13]) which followed a similar idea: lifting the low-level image evidence to some inter-mediate level. Both works first lift the image evidence to a 2D pose estimate and then use this 2D pose estimate to compare it with projected 3D body model configurations. [13] estimate consistent sequences of 2D poses which are called 'tracklets' using the pictorial structures model and use these tracklets as input for the 3D pose estimation. [12] use non-parametric belief propagation (NBP) to infer probability distributions representing the belief in the 2D pose state of each limb. Nevertheless, in both approaches an early decision about the 2D pose is made by using a 2D kinematic model which restricts the space of possible 2D poses and thereby the space of possible 3D poses that can be detected if the 2D kinematic model does not include all possible projections of all 3D poses. In contrast, we do not use any 2D kinematic model and do not even try to estimate a 2D pose since we

believe that 2D pose ambiguities can better be solved in the 3D world. For this, the output of our first stage is a vote distribution for possible marker locations and not a set of possible 2D poses.

This article is structured as follows. In section 2 we explain how we determine potential locations for each 2D body marker, while section 3 describes the 2D and 3D pose estimation step. Section 4 evaluates our approach on the recently published Utrecht Multi-Person Motion (UMPM) benchmark designed especially for quantitative 3D pose estimation evaluations.

2 2D Body Marker Location Estimation

In this section we describe how we transform a set of local image features (we use SURF features) – extracted from a person image – into a set of potential marker locations.

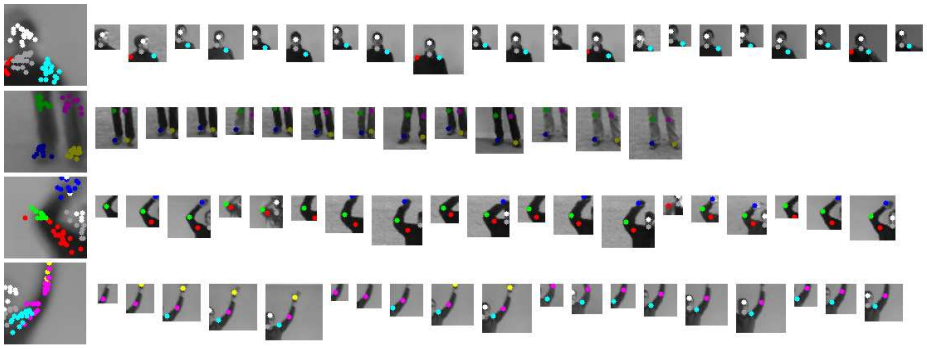

Fig. 1. Small images: sample descriptor regions of extracted SURF features that are assigned to the same visual word. Only the ground-truth marker locations within the corresponding descriptor region are used for training samples of (visual word, marker location) pairs. Ground-truth marker locations as head, neck, shoulders, elbows, hands, knees, and feet are visualized by circles. Big images at left: visual word mean image and corresponding collected marker locations from the samples. The mean image for the visual word is computed by averaging all descriptor region images (resized to fixed mean image size) of the features to the right.

We assume that we already know the approximate location of the persons in the image. State of the art person detectors for this task are sliding detection window based approaches as the HOG detector [14] and part-based approaches as e.g. the Implicit Shape Model (ISM) [10] which uses SIFT features, or the Fastest Pedestrian Detector in the West [15] which uses Integral Channel Features [16]. For a survey on state of the art person detection we refer the reader to [17].

For the body marker location estimation step we adopt the ISM based body part detection approach presented in [11] but introduce three important modifications to the original work. The basic idea of ISM based 2D pose estimation

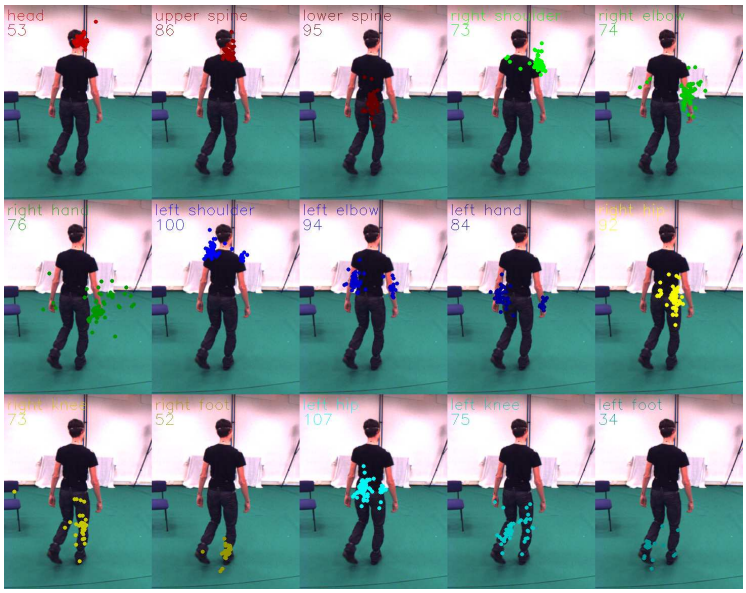

Fig. 2. Body part location ambiguity. For each marker we collect votes casted from visual words. Votes are displayed by small circles. Deciding for an unique location for each marker is not possible at this early stage of pose estimation, since left/right ambiguities cannot be solved just based on identified image structures. Note the left/right ambiguities for nearly all body parts (shoulders, elbows, hands, hips, knees, feet).

is to learn the spatial distribution of body markers relative to each visual word of a codebook. More precisely, we start with a training step in which we first compute local image features for a training set of 2D pose ground-truth labeled person images, map each feature to a visual word of a given codebook (computed by clustering local features extracted from a set of person images), and record the location offsets for each body marker relative to this visual word. The result is a list of example locations for each (visual word, marker) pair. These lists are the training result and will be used later to estimate the 2D pose for a new person image. In the detection phase, visual words then cast marker specific votes according to the previously learned example locations.

A first modification to the original ISM approach concerns the collection of (visual word, marker location) observation pairs. The original approach [10] and its extension to 2D pose estimation [11] used all (feature, marker) observations, even if the marker location was not within the descriptor region of the feature. Thus a local image structure e.g. occurring spatially restricted only near to the feet was later allowed to vote for the right shoulder location as well. We observed this idea to be highly problematic, since votes of some few visual words that e.g. appear on the right shoulder are then typically dominated by a huge set of votes stemming from all other visual words, which actually do not have a meaning for the location of the right shoulder. This leads to the new idea of collecting only

(feature, marker) observation pairs in the training step for which the marker location is within the descriptor region of the corresponding observed feature (see Fig. 1).

The second modification relates the selection of features that are used in the vote casting phase. The SURF keypoint detector typically detects blob-like structures, thus it can provide also a huge amount of meaningless small scale features e.g. on textured blob-like structures appearing on the person's clothings or small image structures. These small region sized features can appear at many different locations on the human body. In the original ISM approach these small sized features are nevertheless matched against visual words and allowed to cast votes, resulting in many wrong votes, which again can dominate the votes casted by visual words covering large parts of the person image and containing worthwhile information about the marker locations. Therefore we discard all extracted SURF features that have a descriptor region size smaller than 15% of the average descriptor region size of all features extracted from the current person image.

A third important modification concerns the final 2D pose estimate determination and representation. In [11] vote maxima were considered to decide for a final location for each marker, thereby making an early decision about the 2D pose although such a decision is mostly not reasonable, since 2D left/right ambiguities and problems due to missing or wrong estimated marker locations can better be solved in the 3D model domain (see Fig. 2). For this, we use the full body marker location vote distributions as input for the determination of a plausible 3D pose.

The ISMs transform the image representation consisting of a set of extracted SURF features to a vote distribution $\mathcal{E}_m$ for each marker which can be considered as a new image evidence $\mathcal{E} = \{\mathcal{E}_m\}$ $(m = 1, \ldots, M)$, where $\mathcal{E}_m = \{\boldsymbol{v_k} = (x_k, y_k)\}$ $(k = 1, \ldots, N_m)$ is the 2D vote distribution for marker m. N_m is the number of votes for marker m in the current frame. M is the number of markers of the body model used (here: $M = 15$).

3 Pose Estimation

This section explains how we estimate a 3D and a 2D pose based on the estimated marker locations, represented by the vote distributions $\mathcal{E}_m$ for each marker.

The pose estimation problem can be formulated using Bayes' theorem:

$$P(\boldsymbol{o}|\mathcal{E}) \propto P(\mathcal{E}|\boldsymbol{o})P(\boldsymbol{o}) \tag{1}$$

where $\boldsymbol{o}$ is a 3D pose. The final 3D pose estimate is then the pose that maximizes the posterior probability $P(\boldsymbol{o}|\mathcal{E})$. Many approaches (e.g. [13], [18], [7]) follow the Bayesian formulation. The difference between the approaches is how the prior probability $P(\boldsymbol{o})$ over the 3D pose space and the observation likelihood $P(\mathcal{E}|\boldsymbol{o})$ are approximated.

Prior Modeling. For modeling the prior $P(\boldsymbol{o})$ we use an example-based approach, i.e. use a set of example 3D poses that have prototypical character.

For this, we traverse a motion capture database (here: UMPM dataset) and incrementally collect example 3D poses. While traversing we consider each 3D pose, compare it to all 3D poses collected so far and only add it to our example set S, if the average joint angle difference is above some threshold θ. This threshold θ allows to control how different the poses in the example set will be: a large threshold e.g. will result in a small example set with large joint angle differences between the example 3D poses. We use the same prior probability $1/|S|$ for each of the 3D example poses, since the occurrence frequency of a 3D pose in a motion capture database does not need to correspond to its occurrence frequency in real world. Example based approaches can only recognize 3D poses seen before and contained in the example set S, which seems to be very restrictive. It depends on the size of S and whether there are poses in S which are similar to the poses we test on, how restrictive this approach is in practice. On the other hand, a major advantage of such an example-based approach is its robustness. We can put similar example 3D poses into S to the ones we expect to be of interest in our application scenario and we do not need to worry about invalid 3D pose configurations as final pose estimates, which is a major issue in geometric reconstruction based approaches as [19].

Likelihood Estimation. The most important issue concerns the question how to model the likelihood $P(\mathcal{E}|\boldsymbol{o})$. The key idea is to project each of our example 3D poses $\boldsymbol{o} \in S$ onto the image and to approximate the likelihood by comparing the projected marker locations with the marker specific votes, i.e. the evidence E_m for potential locations of the m-th marker (see Fig. 3).

Since we do not know from which viewpoint we will observe the person, we generate for each pose $\boldsymbol{o}$ a set of 2D example projections while rotating and tilting the pose – encoded by two angles α_{pan} and α_{tilt} or a single rotation matrix $\boldsymbol{R}$ (where we use only 2 DOFs, no camera-up vector). Different focal lengths f are used to handle different perspective limb foreshortening situations. If we do not expect the person images to be recorded from a strong perspective, we use a large focal length to approximate the parallel projection. Each 3D marker point $q = (x, y, z)$ is projected to a 2D image point $q' = (u, v)$ using a perspective camera model, i.e.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \boldsymbol{R} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + t, \qquad u = f\frac{x'}{z'} + p_x, \qquad v = f\frac{y'}{z'} + p_y \tag{2}$$

To handle different locations of the person in the image, we further use different principal points $\boldsymbol{p} = (p_x, p_y)$ for the final location of the projected pose. If we know the rough person center (e.g. due to a person bounding box) we can use locations $\boldsymbol{p}$ sampled around this person center. To handle different sizes of the person, we further scale the final projected 2D pose limb lengths with an uniform scaling factor s. A single projection of a pose $\boldsymbol{o}$ can therefore be described by the projection parameters $\boldsymbol{c} = (\alpha_{pan}, \alpha_{tilt}, f, \boldsymbol{p}, s)$. For each 3D pose $\boldsymbol{o}$ we generate a set $T(\boldsymbol{o}) = \{\boldsymbol{o'}\}$ of example projections $\boldsymbol{o'} = \{q'_i = (u_i, v_i)\}$ $(i = 1, \ldots, M)$ which we can compare individually with the image evidence $\mathcal{E}$ by comparing

each projected marker location $\boldsymbol{q'_i} = (u_i, v_i)$ with all the votes $\boldsymbol{v_k} \in \mathcal{E}_m$ for this marker.

For this, we compute a weighted sum of all votes which are nearby to the projected marker location, i.e. assess how much evidence we can find that the observed marker location $\mathcal{E}_m$ is near to the projected marker location $\boldsymbol{q'_i}$:

$$P(\mathcal{E}|\boldsymbol{o'}) = \sum_{i=1}^{M} \mathbf{f}(\boldsymbol{q'_i}, \mathcal{E}_i) \tag{3}$$

where $\mathbf{f}$ is some kernel function that weights the votes in $\mathcal{E}_i$ depending on their distance to the projected marker location $\boldsymbol{q'_i}$. The simplest choice for $\mathbf{f}$ is a flat kernel, i.e. we count the number of votes that are within a circle around the projected marker which represents the kernel center. Since the meaning of *nearby* depends on the overall size of the projected pose $\boldsymbol{o'}$, we should adapt the kernel size proportional to the scale factor s, i.e. $r \propto s$. For a flat kernel, $\mathbf{f}$ is defined as:

$$\mathbf{f}(\boldsymbol{q'_i}, \mathcal{E}_i) = \sum_{\boldsymbol{v_k} \in \mathcal{E}_i} \mathbf{g}(\boldsymbol{q'_i}, \boldsymbol{v_k}) \qquad \mathbf{g}(\boldsymbol{q'_i}, \boldsymbol{v_k}) = \begin{cases} 1 \,, \|\boldsymbol{q'_i} - \boldsymbol{v_k}\| \leq r \\ 0 \,, \|\boldsymbol{q'_i} - \boldsymbol{v_k}\| > r \end{cases} \tag{4}$$

While the flat kernel discards all votes having an Euclidean distance greater than r from the projected marker, a Gaussian kernel function $\mathbf{f}$ can be used to consider all votes, but giving the votes different weights depending on their distance to the projected marker location:

$$\mathbf{g}(\boldsymbol{q'_i}, \boldsymbol{v_k}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\|\boldsymbol{q'_i} - \boldsymbol{v_k}\|}{\sigma})^2} \tag{5}$$

where the variance of the Gaussian should be set again according to the scale of the 2D pose, i.e. $\sigma \propto s$.

The overall likelihood for a single 3D pose $\boldsymbol{o}$ is

$$P(\mathcal{E}|\boldsymbol{o}) = \max_{\boldsymbol{o'} \in T(o)} P(\mathcal{E}|\boldsymbol{o'}) \tag{6}$$

Since we find the best 3D pose by comparing projections of our example 3D poses with the image evidence under different projection situations, the final result of the pose estimation step is not only (i) a 3D pose $\boldsymbol{o}$, but (ii) also a 2D pose estimate $\boldsymbol{o'}$ as well, and further (iii) the overall relative orientation of the person to the camera, since we know from which camera view angles α_{pan}, α_{tilt} we projected the 3D pose $\boldsymbol{o}$ to the 2D pose $\boldsymbol{o'}$.

4 Evaluation

In this section we evaluate the quality of our 3D pose estimation method by comparing ground truth 3D poses with our estimated 3D poses.

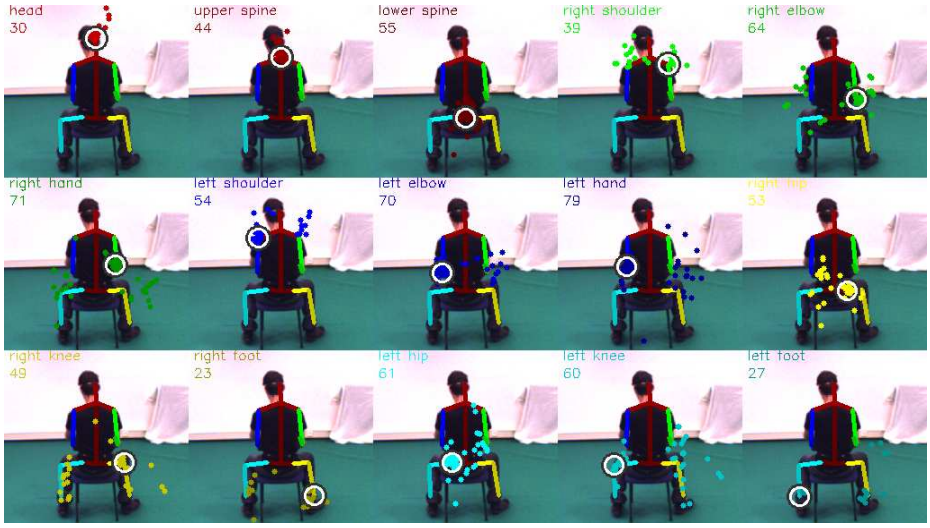

Fig. 3. Observation likelihood computation principle. Example 3D poses $\boldsymbol{o}$ are projected into the image using different projection parameters $\boldsymbol{c}$. For comparing the projected pose $\boldsymbol{o'}$ with the image evidence $\mathcal{E}$ we count how many votes we can find nearby each projected marker $\boldsymbol{q'_m}$ within some radius (visualized by circle). This is done independently for each marker m using the marker specific vote lists $\mathcal{E}_m$. While the flat kernel does consider only nearby votes, the Gaussian weighting strategy considers all votes, but weighted with their distance to the projected marker location $\boldsymbol{q'_m}$. The pose $\boldsymbol{o'}$ displayed is the best match found for the vote distribution in this frame.

Test dataset. The new [20] Utrecht Multi-Person Motion (UMPM) benchmark dataset allows for detailed comparisons of estimated vs. ground-truth poses. It includes synchronized motion capture and video data for 4 cameras and therefore allows to compare an estimated 3D pose with the corresponding ground truth 3D pose. It further provides the extrinsic (camera location + rotation) and intrinsic (focal lengths, principal points, distortion polynomial coefficients) parameters for each of the 4 cameras. Thus we can project each 3D pose into the image to yield the corresponding ground truth 2D pose as well, which we can use for generating training pairs (image, 2D pose) for training the body marker localizer described in section 2.

Error measures. We evaluate our pose estimates quantitatively using two different error measures: E_1 is used to measure the articulation error and is defined as the average joint angle difference between the estimated and the ground-truth 3D pose. E_2 is used to measure the global orientation angle error. For this, we project the line connecting left and right hip onto the floor (XZ plane in the OpenGL visualizations in Fig.4) and define the angle between this line and the camera plane as the global orientation. E_2 measures the relative

`http://www.projects.science.uu.nl/umpm/`

exp no.	experiment description	Train 2D [frames] / [persons]	$\|S\|$ [poses]	Test [frames]	E1 (var) [°]	E2 (var) [°]
1	calibration + flat + limited S	535 / 1	148	514	8.3 (5.2)	29.6 (33.1)
2	calibration + flat + exhaustive S	535 / 1	514	514	7.7 (3.9)	34.7 (45.5)
3	calibration + Gaussian + limited S	535 / 1	148	514	8.4 (4.8)	39.5 (49.2)
4	calibration + Gaussian + exhaustive S	535 / 1	514	514	7.8 (4.3)	36.6 (43.2)
5	generalization + flat + limited S	213 / 4	245	504	9.6 (4.8)	25.6 (2.3)
6	generalization + flat + exhaustive S	213 / 4	504	504	5.5 (3.2)	17.7 (3.4)
7	generalization + Gaussian + limited S	213 / 4	245	504	11.4 (7.1)	20.3 (3.1)
8	generalization + Gaussian + exhaustive S	213 / 4	504	504	5.5 (3.0)	16.5 (2.2)

Table 1. Experiment definitions and results. The average over all 8 experiments for E_1 is 8.0° (36.3°), while for E_2 it is 27.6° (22.8°).

angle difference between the global orientation of the estimated and the ground truth pose.

Limited vs. exhaustive example poses. Since the smallest reachable pose error E_1 is limited by the similarity between our example 3D poses in our example set S and the actual ground-truth 3D poses we conduct two experiments. In the first setting (limited set of example poses), we use a training set S with prototype 3D poses, which were extracted from motion capture sequences of other person, i.e. different to the ones we test on, using the procedure described in section 3 by collecting example 3D poses ($\theta = 0.05$ radians). This is the realistic setting in which we do not have knowledge about the 3D poses that will occur in the test phase. In the second setting (exhaustive set of example poses), we put all 3D poses of the sequence we test on into the training set. This is an unrealistic setting in terms of a later application, since we know which 3D poses will occur before. Nevertheless, it is interesting to see which quality of 3D pose estimates we can reach if S is perfect, since (i) the body part localization step and (ii) the discrete sampling of projection configurations $\boldsymbol{c}$ and the (iii) search for the right 3D pose within the set of examples poses will introduce additional errors into the overall processing pipeline.

Calibration vs. Generalization. A further important distinction has to be made between scenarios, (i) in which we train the ISM body part localizer on a person P_1 and test the pose estimation on new sequences showing P_1 (experiments 1-4), or (ii) in which we train the ISM body part localizer on some persons $P_1, \ldots, P_N$ and test on a new person P_{N+1} which was never seen before (experiments 5-8). In (i) we will have very similar local features in the test phase as in the training phase, whereas in (ii) train and test local features will differ and therefore the generalization ability of the SURF feature descriptor, the feature to visual word matching, and the ability of the successive pose estimator to compensate for wrong estimated body marker locations is of high importance. Since (i) is nevertheless an important application scenario in which we calibrate the 2D pose estimator on a person using a marker-based approach and can later estimate poses marker-less using monocular camera information only for this person, we test both scenarios.

Flat vs. Gaussian kernel. We further differ between experiments in which we used the flat kernel vs. the Gaussian kernel (see section 3).

Experiments conducted. Table 1 shows the experiments conducted and the results regarding E_1 and E_2 (mean and variance of these errors). We further specify the total number of training and test frames used. In experiment 5 e.g. we train the ISM based body part localizer using 213 (image, 2D ground truth pose) pairs of 4 different persons $P_1, \ldots, P_4$, while the set of example poses S contains 245 example 3D poses (from motion capture data from persons $P_1, \ldots, P_4$). Testing (3D pose estimation) is done for a sequence of 504 frames showing a new person P_5 never seen before (image, motion capture data), resulting in an average joint angle error (averaged over all joint angles and 504 frames) of 9.6° (variance of this error: 4.8°), while the global orientation error was 25.6°.

Computing time. For projecting a single 3D pose onto the 2D image and comparing it with the image evidence using the flat kernel only 0.00222 ms = 2.22 μs are needed in average (straightforward C++ implementation, standard hardware), i.e. we can roughly project and compare 450450 3D pose configurations with the body marker vote distributions per second. For the Gaussian vote weighting strategy we need in average 0.0035 ms = 3.5 μs which allows to test 285714 3D pose configurations per second.

Results. The results show that using an exhaustive example 3D pose set S (exps 2,4,6,8) yields better results compared to the limited example 3D pose set (exps 1,3,5,7), which we expected. Nevertheless, there is a huge difference between the calibration and the generalization scenario. For the calibration scenario there was no significant difference, while for the generalization scenario, the average joint angle error dropped e.g. from 9.6° to 5.5° (exp 5 vs. 6), and from 11.4° to 5.5° (exp 7 vs. 8) for E_1. This could indicate that considering as many 3D poses as possible can dramatically help to improve 3D pose estimation, which in turn underlines the need for a lightweight processing pipeline for generative approaches as presented here, in order to allow for testing several thousands of different 3D poses and configurations per frame. The results for the flat vs. Gaussian kernel were unexpected. We expected the Gaussian kernel to yield better results than the flat kernel since it weights votes nearer to the projected marker locations higher than far distant votes and thereby distinguishing between near and far distant votes. But the opposite was true, which is important to know, since the usage of an Gaussian (exponential) kernel is connected with higher computational costs (3.5 μs vs. 2.22 μs per pose evaluation): for the Gaussian kernel we need to evaluate the exponential function, while for the flat kernel, we only need to consider the distance of a vote to the projected marker location. The average joint angle error – averaged over all joints, frames, and experiments – is 8.0° (E_1), while the average global orientation error for the estimated 3D poses is 27.6° (E_2). The smallest 3D pose articulation error E_1 was reported by Agarwal and Triggs [21] and is 5.91° for single-frame pose estimation and 4.1° for multiple-frame pose estimation. While our average joint angle error is larger, one has to underline, that the results presented in [21] are for artificial 2D training and testing silhouettes, generated using motion capture data.

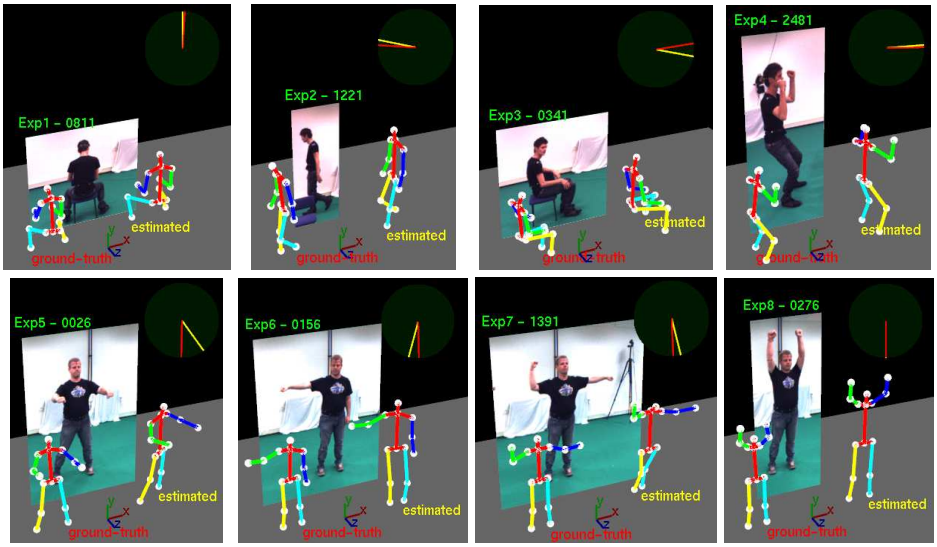

Fig. 4. Qualitative examples of estimated 3D poses. For each of the 8 experiments we show one sample screen-shot of ground-truth and estimated 3D pose (automatically generated from our pose error evaluation tool). Experiment and frame number are rendered in top left corner of the corresponding camera frame image. In each image the left pose is ground-truth, while the right pose is the estimated pose. In the top right corner we visualize the global orientation of the ground-truth and estimated pose. E_2 is the relative angle between the two lines (averaged over all test frames of the corresponding experiment).

5 Conclusions

We have presented a new idea for generative human pose estimation that is quite simplistic, but showed to be effective for recovering the 3D pose of a person. Instead of using low-level image features as edges, or silhouettes on the one hand, or high-level 2D pose estimates on the other hand, for comparing with model projections, we propose to use body marker vote distributions from a ISM based body marker localizer. These can be compared with negligible computational costs with a simple 3D stick-figure body model projected to the image plane. A flat kernel, that counts for the number of votes nearby to projected marker locations of this 3D stick-figure, is sufficient as basis for an observation likelihood that allows to filter for the correct projection parameters and 3D example pose from a set of prototype poses. Additionally, we motivated several modification ideas for a recently published ISM based body part localization approach for improving 2D body marker localization.

Our next step will be to incorporate temporal information into this approach.

References

1. Sminchisescu, C.: 3D human Motion Reconstruction in Monocular Video. Techniques and Challenges. Volume 36 of Human Motion Capture: Modeling, Analysis, Animation. Springer (2007) ISBN 978-1-4020-6692-4.
2. Poppe, R.: Vision-based human motion analysis: An overview. CVIU **108** (2007) 4–18
3. Ji, X., Liu, H.: Advances in view-invariant human motion analysis: A review. IEEE Transactions on Systems, Man, and Cybernetics, Part C **40** (2010) 13–24
4. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. Volume 0., Los Alamitos, CA, USA, IEEE Computer Society (1998) 8
5. Roth, S., Sigal, L., Black, M.J.: Gibbs likelihoods for bayesian tracking. In: CVPR. (2004) 886–893
6. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. CVPR **1** (2003) 69
7. Sigal, L., Balan, A.O., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS. (2007)
8. Drummond, T., Cipolla, R.: Real-time tracking of highly articulated structures in the presence of noisy measurements. In: ICCV. (2001) 315–320
9. Charles, J., Everingham, M.: Learning shape models for monocular human pose estimation from the microsoft xbox kinect. In: ICCV Workshops, IEEE (2011) 1202–1208
10. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. IJCV **77** (2008) 259–289
11. Müller, J., Arens, M.: Human pose estimation with implicit shape models. In: ACM Artemis. ARTEMIS '10, New York, NY, USA, ACM (2010) 9–14
12. Sigal, L., Black, M.J.: Predicting 3d people from 2d pictures. In: AMDO. (2006) 185–195
13. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: Proc. of CVPR 2010, USA (2010)
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. Volume 1. (2005) 886–893
15. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: BMVC, Aberystwyth, UK (2010)
16. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC. (2009)
17. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. PAMI **99** (2011)
18. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009) 1014–1021
19. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. CVIU **80** (2000) 349–363
20. Aa, N.v.d., Luo, X., Giezeman, G., Tan, R., Veltkamp, R.: Utrecht multi-person motion (umpm) benchmark: a multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In: HICV workshop, in conj. with ICCV. (2011)
21. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. PAMI **28** (2006) 44–58