Human Pose Estimation in Stereo Images

Joe Lallemand[1,2], Magdalena Szczot[1], and Slobodan Ilic[2]

[1] BMW Group, Munich, Germany
{joe.lallemand, magdalena.szczot}@bmw.de http://www.bmw.de
[2] Computer Aided Medical Procedures, Technische Universität München, Germany
slobodan.ilic@in.tum.de http://campar.in.tum.de

Abstract. *In this paper, we address the problem of 3D human body pose estimation from depth images acquired by a stereo camera. Compared to the Kinect sensor, stereo cameras work outdoors having a much higher operational range, but produce noisier data. In order to deal with such data, we propose a framework for 3D human pose estimation that relies on random forests. The first contribution is a novel grid-based shape descriptor robust to noisy stereo data that can be used by any classifier. The second contribution is a two step classification procedure, first classifying the body orientation, then proceeding with determining the full 3D pose within this orientation cluster. To validate our method, we introduce a dataset recorded with a stereo camera synchronized with an optical motion capture system that provides ground truth human body poses.*

Keywords: Human Pose Estimation, Machine Learning, Depth Data

1 Introduction

Human body pose estimation in depth images has seen tremendous progress in the last few years. The introduction of Kinect and other similar devices has resulted in a number of new algorithms addressing the problem of 3D human body pose estimation [1–4].

Although these Kinect-like sensors work in real-time and usually provide depth images of a good quality with a small amount of noise and depth errors as depicted in Fig.1, they also have the disadvantages of only working indoors and at a very limited depth range. For these reasons, human pose estimation using Kinect has extensively been used for generic indoor scenarios. Many other applications however, especially automotive driver assistance, imply the use of outdoor-suitable sensors as e.g. stereo cameras. Since stereo camera systems are becoming standard in modern cars, there is a need for 3D human pose estimation from stereo data. For that reason, we propose a new algorithm using a stereo camera which provides real time depth images at a range of up to 50 meters, which is about 5 times higher than indoor sensors. As can be seen in Fig.1, real-time stereo algorithms integrated in vehicles generally produce noisy images, where some regions are erroneously fused together (red circles) and the boundaries of the objects can be effected by a high number of artifacts (green and blue

circles). These reconstruction artifacts introduced by stereo reconstruction affect the results of state-of-the-art methods, like the one of Grishick et al. [2], which we reimplemented and applied to these data. This is because of the two main reasons. Firstly, it is very difficult to perform 360 degree pose estimation using a single forest as there is a high confusion between front and back. Secondly the feature vector proposed in [1] seems to perform poorly on the stereo data. Therefore we present a new method for human pose estimation, adapting random forest classification and regression methodology into a two step pipeline to reliably estimate 3D human body pose. The first step consists in classifying the shape of a person into a cluster which represents its orientation with respect to the camera. In the second step, the skeleton pose of the person is estimated using a regression random forest trained only on the poses of the detected orientation. In order to make this pipeline operational we introduce a novel grid-based feature. This feature overcomes several disadvantages that appear when using the depth comparison feature introduced by Shotton et al. [1] on the stereo data as shown in the result section.

To verify and validate our method, we introduce a dataset which is recorded with a stereo camera synchronized with the ART marker based system for human motion capture [3]. The orientation classification is also evaluated on the publicly available pedestrian stereo data set introduced in [5].

2 Related Work

Many algorithms for human pose estimation from depth images have emerged in the last years. Shotton et al. [1] propose to use a classification random forest to classify each pixel of a foreground mask to a given body part, then infer the joint locations from the predicted body parts. Girshick et al. [2] extended this work by learning a regression model to directly predict the joint locations. This approach considerably improved the performance of the previous algorithm especially for occluded joints. Both works rely on a large synthetic training dataset in order to achieve good results and target good quality depth images.

In [3], Taylor et al. train a regression random forest to create a mapping from each pixel of a segmented foreground depth image to a human body model. Taking into account the forest predictions, physical constraints and visibility constraints, they use an energy minimization function to predict the pose of the model and the attached skeleton. This approach improves prediction accuracy compared to previous works and is able to predict poses in the 360 degree range, but still relies on the tree structures trained using the classification approach of [1].

Sun et al. [6] introduce a conditional regression forest learning a global latent variable during the forest training step that incorporates dependency relationships of the output variables, e.g. torso orientation or height.

A simple depth comparison feature is common to all these methods. Each dimension of it consists of the difference in depth computed at two random offsets

[3] www.ar-tracking.com

from the reference pixel at which the feature is computed. As the foreground masks in stereo data contain many erroneous boundaries, the feature cannot be consistently extracted for the same pose. The proposed grid-based feature is robust to these errors because it consists of cells where depth and occupancy distribution are averaged over the whole cell.

Plänkers and Fua [7] use an articulated soft object model to describe the human body and track it in a system of calibrated video cameras, making use of stereo and silhouette data. Urtasun and Fua [8] additionally introduce a temporal motion models based on Principal Component Analysis. Bernier et al. [9] propose a 3D body tracking algorithm on stereo data. The human body is represented using a graphical model and tracking is performed using non-parametric belief propagation to get a frame by frame pose. Unlike the three previously mentioned works, which require initialization and track the human pose, our proposed method works on single frames and performs discriminative pose estimation. Up to the best of our knowledge, this problem has not yet been addressed for the kind of noisy input data as produced by stereo cameras or similar devices.

Keskin et al. [10] use a two-layer random forest for hand pose estimation. First, the hand is classified based on the shape, then the skeleton is determined for the given shape cluster using a regression forest. Though similar to [10], we introduce a novel grid-based feature and a two stage classification method for human pose estimation in noisy stereo data.

In [11], Enzweiler and Gavrila propose an algorithm for single-frame pedestrian detection and orientation estimation based on a monocular camera, where orientation is divided into 4 directions. In contrast to this, the proposed method is based on the depth information from the stereo camera and the orientation clusters are encoding direction as well as different poses within this direction.

3 Method

This section introduces the grid-based feature vector which is used both, for the classification of human body orientations and the human pose estimation per determined orientation and describes the two step classification pipeline. The first step involves determining the human body orientation. While the second computes the 3D pose of a skeleton choosing from poses of the estimated orientation cluster. Finally, we describe how the classification and pose prediction step are combined.

3.1 Grid-based Feature

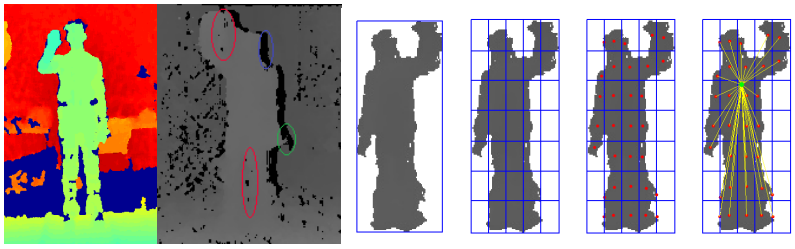
The proposed grid-based feature divides the shape of a person into arbitrary cells, then averages over depth values and occupancy distributions.

Let $\Omega \subset \mathbb{R}^2$ be a segmented foreground mask in a given image. The construction of the feature vector consist of 4 consecutive steps. The first step determines the bounding box around the foreground mask. In the second step, the bounding

box is divided into an $n \times m$ grid of cells $c_{i,j}$. Note that this division is scale invariant, as the bounding box, regardless of its actual size, is divided into the same number of cells. In the third step, we attribute each pixel of the foreground to its corresponding cell and determine the median position, $x_{c_{i,j}} \in \mathbb{R}$ and $y_{c_{i,j}} \in \mathbb{R}$ and median depth $z_{c_{i,j}} \in \mathbb{R}$ in each cell. This cell structure now represents a very simple encoding of the shape of a person. If a cell is left unoccupied, it is assigned a very high value. Finally, the pixel-wise grid-based feature is given by:

$$f_{p_k} = \{x_k - x_{c_{1,1}}, y_k - y_{c_{1,1}}, z_k - z_{c_{1,1}}, \ldots, x_k - x_{c_{n,m}}, y_k - y_{c_{n,m}}, z_k - z_{c_{n,m}}\} \quad (1)$$

for a pixel $p_k = \{x_k, y_k, z_k\}$. Figure 1 shows the different steps of generating the feature vector. In this way, the feature vector is able to ignore small errors of the stereo algorithm especially around borders and systematic errors of the algorithm are taken into consideration as shown in Fig. 1 (b). The result section provides analysis of the influence of the feature dimension on the performance of the classifier.



(a) (b) (c)

Fig. 1: (a,b): Comparison between the data quality acquired with Kinect(a) and with the stereo camera(b). (c) Different stages of creating the feature vector here for 5×7 cells from left to right: the bounding box tightly laid around the foreground mask, the subdivision of the bounding box into a grid of cells, the computed median in each cell in red, the feature vector for a randomly chosen pixel in green and the connection to each cell median in yellow.

3.2 General Theory on Training Random Forests

A random forest is an ensemble of decorrelated binary decision trees. It is trained on a dataset Δ, consisting of pairs $\psi_i = \{f_i, l_i\}$ of feature vectors f and the labels l , learning the mapping from the features to the labels. Each tree is trained on a subset of the training data ensuring that the trees are randomized. At each node of the tree, a decision function $g_{\nu,\tau}(f) \equiv \nu * f < \tau$ is trained sending samples to the left child node if this condition is verified else to the right child node, where ν chooses exactly one feature dimension thus creating axis aligned splits. In order

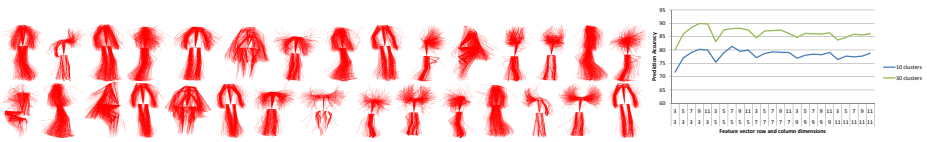
to train this decision function, at each node, a subset of all feature dimensions is randomly chosen and for each feature, n thresholds are generated, separating the incoming samples Δ into left and right subsets Δ_l and Δ_r. For each of these splits, an information gain is computed:

$$\mathbf{I}_{\nu,\tau} = -\frac{|\Delta_l|}{|\Delta|} H(\Delta_l) - \frac{|\Delta_r|}{|\Delta|} H(\Delta_r) \tag{2}$$

where H is an entropy function depending on the kind of random forest and $|\cdot|$ denotes the number of elements in a set. The final decision function g_{ν^*,τ^*} is given by finding $argmax_{\nu,\tau}(I_{\nu,\tau})$. This process is repeated iteratively until a leaf node is reached, which is defined by the following criteria: (i) the maximum depth is reached, (ii) a minimum number of samples is undercut or (iii) the information gain falls below a certain threshold. In the leaf nodes, all incoming samples are used to compute a posterior distribution which depends directly on the kind of forest trained.

3.3 Orientation Classification

The goal of the orientation classification is to assign the current foreground mask to its corresponding cluster containing all the poses of a specific orientation in relation to the camera. To achieve this, clusters are created using the motion capture data acquired for each pose and a classification random forest is trained to classify each pixel into the correct cluster.



(a) (b)

Fig. 2: (a)30 Orientation clusters obtained with k-means clustering. For such a large number of clusters, the poses are divided by orientation but also broadly into arm and leg movements. (b) Orientation classification results for different sizes of the grid like feature and different number of orientation cluster.

Generation of Orientation Clusters. The clusters are generated in an unsupervised manner, using the motion capture data from the training dataset. For each pose, the angles between all neighboring joints are computed. Clustering is done using the k-means approach on these joint angles. In case k-means is run on the euclidean distances of joint positions in 3D space, the algorithm not only separates poses in terms of joint angles but also people of different heights. By using

only the joint angles and deliberately omitting limb lengths, we get consistent clusters for different poses with regard to the overall orientation of the person. K-means relies on initial seeds to create clusters and results can vary depending on those seeds. In order to achieve a certain level of independence from this, we run 100 instances of k-means and choose a cluster combination which is most often reached during this process. The influence of the number of clusters is analyzed in the Sec 4. Although other clustering algorithms, e.g. mean shift [12] were tested, they didn't give satisfactory results. Since fixing the bandwidth of mean shift by hand is not trivial. K-means was the final choice for clustering.

Classification of Orientation Clusters. The classification random forest is trained using the grid-based feature to classify each pixel to the correct cluster. Shannon's entropy is used for the information gain. Additionally, we use the random field based reweighting function described in [13]. This reweighting scheme takes into account the class distribution of the full training dataset, instead of reweighting only the samples in the current node, which was shown to yield more accurate results. The information gain I is rewritten as:

$$I_{\nu,\tau} = - \sum_{i \in \{l,r\}} Z\left(\Delta_i\right) \sum_{c \in C} n\left(c, \Delta_i\right) \log \left(\frac{w_c n\left(c, \Delta_i\right)}{Z\left(\Delta_i\right)} \right) \tag{3}$$

where Δ_0 is the total training set, $n\left(c, \Delta_i\right)$ is the number of occurrences of class c in the subset Δ_i, and $w_c = \frac{\sum_{k \in C} n(k, \Delta_0)}{n(c, \Delta_0)}$ is the weight obtained by dividing the total number of samples k in the dataset Δ_0 by the number of occurrences of class c. It is lowest for the most represented class and vice versa. $Z\left(\Delta_i\right) = \sum_{k \in C} w_k n\left(k, \Delta_i\right)$ is analogous to the partition function in a random field and represents the weight of a given subset Δ_i. It replaces the weight $\frac{|\Delta_l|}{|\Delta|}$ in Equation 2. The detailed derivation of this formula from the standard Shannon's entropy is presented in the works of Kontschieder et al. [13] where this new information gain was first introduced. The leaf nodes store the distribution of classes of all incoming points as a histogram.

3.4 Pose Estimation per Orientation Cluster

One regression forest is trained for the pose estimation of each cluster. For each tree, the training set consists of pixels obtained from a bootstrap of the training images belonging to a given cluster. The ground truth joint positions are provided by a motion capture system, as will be explained in Sec. 4.1. The training dataset consists of pairs of pixel-wise features as described in Sec. 3.1 and labels containing the offset from the given pixel to all joint positions. For a given pixel $p_i(x_i, y_i, z_i)$ and the pose $J = \{j_1, \ldots, j_N\}$ consisting of N joints, the label is given by $\Psi = \{\psi_1, \ldots, \psi_N\}$, with each $\psi_k = (j_{k,x} - x_i, j_{k,y} - y_i, j_{k,z} - z_i)$. During training we iteratively search for the optimal split in each node. As shown in [2], the body joints can be modeled using a multivariate gaussian distribution. Following this idea, we can model the information gain based on the differential

entropy of gaussians and assume independence between the different joints. The entropy function H in the information gain function can thus be reduced to:

$$H\left(\Delta\right) = \frac{1}{2}\log\left(\left(2\pi e\right)^{3N}\left|\Sigma^{(\Delta)}\right|\right) \tag{4}$$

where Σ is the diagonal of the covariance matrix of the joint positions and N is the number of joints. Once a leaf node criterion is fulfilled, the mean shift is computed on all incoming points for each joint and the main mode is stored with its weight, equal to the number of points voting for the main mode .

3.5 Prediction Pipeline

For each image with an unknown pose, the grid-based feature is computed for a random subset of pixels from the foreground mask. They are then sent through all trees of the orientation classification forest. The histograms, containing the distribution over orientation clusters, extracted from all leafs are averaged over all pixels and trees. We retain the three best orientations for the pose estimation. In the pose estimation step, all pixels are sent through the forests belonging to those three best orientation clusters. The final pose aggregation is done by applying mean shift to the predictions for each joint separately and choosing the main mode as the prediction outcome.

4 Experiments and Results

4.1 Data Acquisition

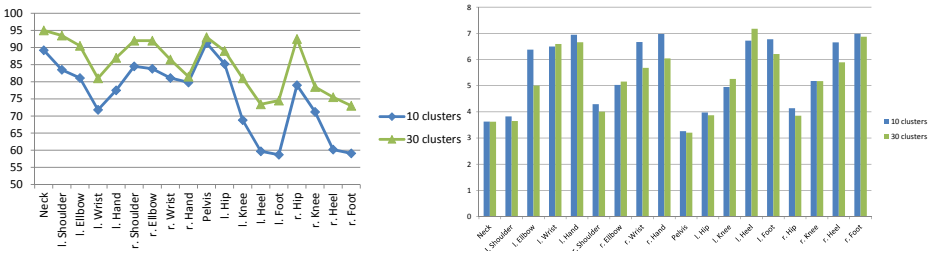
In order to be able to test our algorithm, we have created a new dataset, using a stereo camera and a motion capture system. Since the mocap system does not work outdoors, the training data was acquired indoors. The training set consists of sequences depicting 10 people performing various walking and arm movement motions. During the acquisition the actors were wearing 14 markers which reflect the infrared light emitted by 8 infrared cameras and are used to provide ground truth skeleton positions for each frame. The dataset consists of 25000 frames.

4.2 Orientation Classification

Proposed Dataset: In this paragraph, we analyze the orientation classification part, described in Section 3.3. The evaluation is twofold, first we analyze how the number of clusters affects the classification outcome, then we evaluate the influence of the number of cells of the feature vector and compare to the depth comparison feature. The number of clusters were set to 10 and 30 during the experiments. For the feature vector, we perform an evaluation progressively increasing the number of cells from 3×3 to 11×11 in steps of 2. The maximum allowed tree depth is set to 20, and each forest consists of 5 trees. All results are averaged over a cross validation. For each validation, the forests were trained on

8 people and tested on the remaining 2. Results can be seen in Fig.2 (b). The best results are achieved for 30 clusters. There are two important observations regarding the feature vector. Firstly, dividing the feature into too many cells, especially along the y-axis, decreases the overall performance of the feature. Especially for side views and poses where all limbs are close to the body, a fine grid along the y-axis negatively effects the noise reduction properties for which the feature was designed. Secondly, the feature vector seems to perform best if the ratio between the number of rows and columns is closer to the height versus width ratio of the human body.

In order to compare the grid-based feature to the feature used in [1–3], we trained a random forest sampling 500 feature candidates and 20 thresholds at each node with a maximum probe offset of 1.29 pixel-meters, identical to those proposed in [1]. All other parameters were kept identical to the other experiments. The grid-based feature achieved 81.4% and 89.9% for 10 and 30 clusters respectively compared to 64.6% and 72.3% for the depth comparison feature used in [1].



(a) (b)

Fig. 3: Evaluation of the grid-based feature vector with regard to the number of clusters and the number of cells in the grid. (a): The accuracy per joint (b):error in cm per joint.

Daimler Pedestrian Segmentation Benchmark: In order to show that the approach also works outdoors, we evaluate the orientation classification on the publicly available dataset of Flohr and Gavrila [5], consisting of 785 single disparity images of pedestrians at various distances from the camera. This dataset contains annotated groundtruth for the foreground masks of the pedestrians but does not contain orientation information. To evaluate our approach, we separate the orientation clusters of our approach into 8 directions with regard to the camera {front, front-left, left, back-left, back, back-right, right and front-right}, choosing for each of the generated clusters the dominant torso orientation. Since the ground truth pose is not available for this dataset to determine the correct cluster, we choose visually the closest orientation used for the manually labeled

clusters. Tests were run for the 30-cluster training setup using the best feature from the previous experiments, achieving 78% accuracy.

It is noteworthy that most of the disparity images provided by the dataset are much smaller in size than the training images. In only about half of the provided images, the height of the foreground mask is higher than 120 pixels, which is roughly half of the average height of the training images. This shows that our algorithm and especially the feature work well even if the size of testing images is a fraction of the size of the training images In Fig. 4, we show some example images from the dataset with the determined orientation.

Fig. 4: Example images from the dataset of [5]. The ground truth label is denoted in green and the prediction in red. The yellow number displays the percentage of foreground pixels voting for the predicted cluster. We show the original image instead of the depth image, as it is visually more helpful.

4.3 Pose Estimation

The evaluation of the pose estimation is done for cluster sizes of 10 and 30. For each scenario, we use the best feature from the previous evaluation and apply the complete prediction pipeline as described in Section 3.5. First the classification forest determines the correct orientation cluster, then the regression forests from the three most probable clusters are used to predict the pose. We consider a joint to be correctly estimated if it is within a radius of 10 centimeters of the ground truth joint position. This follows the evaluation criteria established by several related works [1, 2]. Results are shown in Fig.3 (a). Fig.3 (b) shows the median error per joint. We explicitly use the median, as an error in the orientation classification is propagated to the pose estimation producing wrong poses with per joint errors of up to 1m. By displaying the median error, we can show that if the correct orientation has been determined, the pose prediction produces good results for all different orientations. Examples are shown in supplementary materials video. To compare our grid-based feature to the depth comparison feature of [1], we train regression forests for each cluster using the same parameters as have been described for the orientation classification. For a fair comparison between both features in terms of pose regression, we use the output of the classification forest trained with the grid-based feature. This way, we do not penalize errors of the depth comparison feature in the orientation classification step. The grid-based feature achieved 75.8% and 84.9% for 10 and 30 clusters, compared to 71.3% and 80.0% for the depth comparison feature.

The prediction pipeline including feature computation, orientation classification and the pose prediction run in real-time at 35 fps on an Intel(R) Core(TM) i5-2540 CPU.

5 Conclusion

We propose a new algorithm for human pose estimation in stereo images consisting of two stages procedure, where we first classify global orientation and then predict the pose. We introduced a new grid-based feature vector and proved its effectiveness compared to the commonly used depth comparison feature of [1]. This feature is also used in our two-stage procedure where first a classification forest was used for orientation prediction and then a regression forest is used for pose estimation. In the future, we want to include the color information provided by the stereo camera and consider temporal information to cope with isolated wrong predictions.

References

1. Shotton, J., Fitzgibbon, A., Cook, M., Finocchio, T.S.M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR. (2011)
2. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: ICCV. (2011)
3. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 103–110
4. Pons-Moll, G., Taylor, J., Shotton, J., Hertzmann, A., Fitzgibbon, A.: Metric regression forests for human pose estimation. (2013)
5. Flohr, F., Gavrila, D.M.: Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues
6. Sun, M., Kohli, P., Shotton, J.: Conditional regression forests for human pose estimation. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 3394–3401
7. Plänkers, R., Fua, P.: Articulated soft objects for multi-view shape and motion capture. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10) (2003)
8. Urtasun, R., Fua, P.: 3d human body tracking using deterministic temporal motion models. In: Computer Vision-ECCV 2004. Springer (2004) 92–106
9. Bernier, O., Cheung-Mon-Chan, P., Bouguet, A.: Fast nonparametric belief propagation for real-time stereo articulated body tracking. Computer Vision and Image Understanding **113**(1) (2009) 29–47
10. Keskin, C., Kıraç, F., Kara, Y., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. Computer Vision–ECCV 2012 (2012) 852–863
11. Enzweiler, M., Gavrila, D.M.: Integrated pedestrian classification and orientation estimation. In: Computer Vision and Pattern Recognition (CVPR), IEEE (2010)
12. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5) (2002)
13. Kontschieder, P., Kohli, P., Shotton, J., Criminisi, A.: Geof: Geodesic forests for learning coupled predictors