Leveraging orientation knowledge to enhance human pose estimation methods[⋆]

S. Azrour, S. Piérard, and M. Van Droogenbroeck

INTELSIG Laboratory, Department of Electrical Engineering and Computer Science,
University of Liège, Belgium

Abstract. Predicting accurately and in real-time 3D body joint positions from a depth image is the cornerstone for many safety, biomedical, and entertainment applications. Despite the high quality of the depth images, the accuracy of existing human pose estimation methods from single depth images remains insufficient for some applications. In order to enhance the accuracy, we suggest to leverage a rough orientation estimation to dynamically select a 3D joint position prediction model specialized for this orientation. This orientation estimation can be obtained in real-time either from the image itself, or from any other clue like tracking. We demonstrate the merits of this general principle on a pose estimation method similar to the one used with *Kinect* cameras. Our results show that the accuracy is improved by up to 45.1 %, with respect to a method using the same model for all orientations.

1 Introduction

Markerless pose estimation has attracted much interest since the release of low-cost depth cameras like the *Microsoft Kinect*. Shotton *et al.* and Girshick *et al.* made an important step by presenting methods that infer a full-body pose reconstruction in real-time. Their details were explained, chronologically, in [8], [3], and [9]. Despite this technological breakthrough, the accuracy of human pose estimation from single depth images remains insufficient for some applications.

The straightforward strategy to improve the pose estimation is to substantially increase the size and the diversity of the learning set, but this is costly, impractical, and often impossible. Other ideas to improve the method of Shotton *et al.* have also been developed. Yeung *et al.* [11] presented a way to combine the predictions of two *Kinect* cameras in order to reduce the problems related to unwanted joints positions vibration and bone-length variation observed with the method described in [9]. Wei *et al.* [10] used a method equivalent to [8] in combination with a tracking algorithm and showed that it improved the robustness and the accuracy on the estimation of the joint positions. In this paper, we present a principle for improvement that can be used with any markerless pose estimation method based on machine learning techniques. Instead of taking

[⋆] The final publication is available at Springer's website via http://dx.doi.org/10.1007/978-3-319-41778-3_8.

advantage of additional cameras or filtering the predictions in a post-processing step, we start by estimating the orientation of the observed person.

Our contribution is to show how an estimation of the orientation of the observed person improves the accuracy of a pose estimation algorithm. Our idea consists in slicing the full orientation range into smaller ranges and learning a different model for each of these smaller ranges. When the models are used to recover the pose, given the estimation of the orientation of the observed person, we use the appropriate model to make the predictions for the joints positions. To take into account the uncertainty on the orientation estimation, we consider slightly overlapping orientation ranges when the models are learned. An illustration of our method is shown in Figure 1.

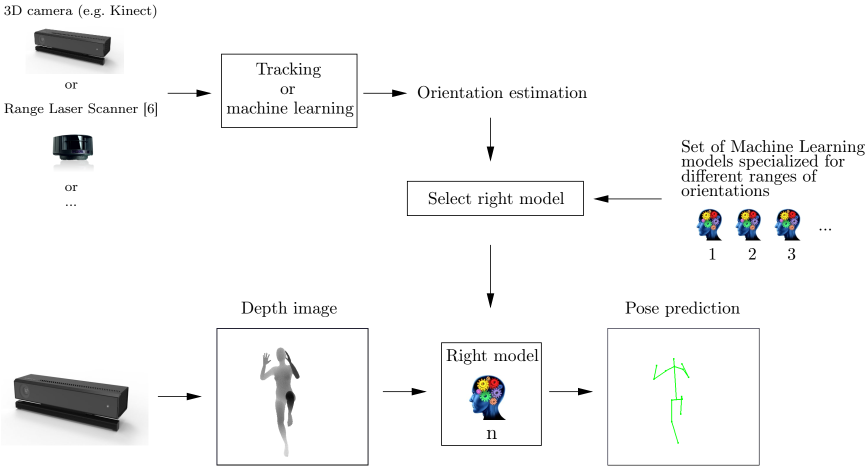

Fig. 1. Outline of our method. The orientation estimation can be obtained from the image itself or thanks to any kind of sensors through a machine learning or a tracking algorithm for instance. The last image shows the skeleton linking the estimated joints.

2 Principle of leveraging an orientation estimation

The intuition for having several models depending on smaller orientation ranges is the following. From our experience, when it comes to analyze silhouettes annotated with depth in each pixel (see Figure 3), machine learning methods tend to grant a high importance to the information related to the external contour and not enough importance to the information related to the depth signal. The problem is that there are two different poses corresponding to the same silhouette shape [7] (when the small details of the silhouette corresponding to the perspective effects are neglected), and this ambiguity leads to large errors when an average solution is predicted. Note that with the arbitrary convention taken in this paper (see Figure 2), one of the two possible poses is associated with an orientation of θ, while the other one is associated with an orientation of $360° - \theta$.

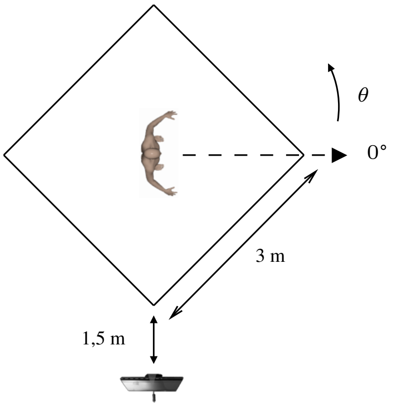

Fig. 2. The configuration considered in this paper. The person can be anywhere in the area (a square of 3 m side) with any pose and any orientation. The camera is placed 1 m above the floor; its optical axis is horizontal.

Therefore, except for the rare cases where the observed person has an orientation very close to 0° (seen from his right side) or 180° (seen from his left side), the knowledge of the orientation is sufficient to overcome the pose ambiguity, even if it is only roughly estimated. Our method is based on the idea that it is preferable to rely on an additional method that is specifically designed for orientation estimation instead of trying to recover the joint positions and disambiguate the silhouette orientation all at once. We observed that when a machine learning method does not have to simultaneously estimate the orientation and the pose, and can focus on the pose estimation given that a rough orientation estimation is provided to it, its task is eased and the accuracy of the predictions is improved.

Several clues can be used to estimate the orientation. When the observed person is walking, his orientation is given by his velocity vector, and can therefore be estimated by tracking. This tracking can be done directly from the depth camera, or from range laser scanners [6]. The orientation can also be estimated directly from a single depth image [5].

One way of forcing the pose estimation method to take the orientation into account is to consider several ranges of orientation and to learn a different model for each range. During the pose estimation step, given the orientation estimation, we use the appropriate model to predict the pose. Note that the overlap between consecutive ranges should be adapted to the maximum uncertainty of the selected orientation estimation method. In the case of the estimation from the depth image, Piérard *et al.* [5] showed that it is possible to achieve an average uncertainty of 4.3° (measured on synthetic, noise-free data), but no bound

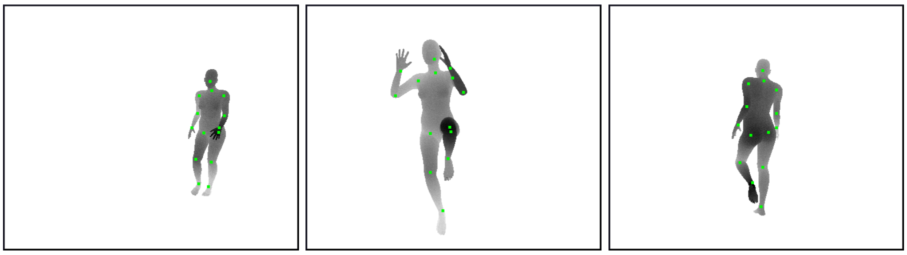

Fig. 3. Examples of generated depth images used in our experiments. The ground truth body joints are displayed in green.

was given. In practice, the errors are larger, but the temporal variance can be filtered out, leading to reliable estimates as shown on the video on the author's website. We take an overlap of 20° for this orientation estimation method.

3 Experiments

To assess the effectiveness of our principle, we implement a simplified (for practical reasons) version of the pose estimation method described in Girshick *et al.* [3]. The main differences are that we use a general regression random forest model (the *ExtRaTrees* [2]) instead of a custom one, that we use 500 features rather than 2,000 to describe the pixels environments, and that the models are learned from another, smaller, dataset.

To generate the learning and test datasets, we followed a method similar to the one described in [9] except that we used the open source softwares *Blender* and *MakeHuman*. Moreover, we used only one human model and did not add clothes to it. Without loss of generality, our small dataset is sufficient to establish that our principle helps to improve the accuracy of the pose estimation. The poses used to generate the data were taken randomly from the *CMU motion capture database* [1]. A few unrealistic poses, that do not correspond to a standing person, have been manually excluded (less than 1%). A total of 24,000 silhouettes annotated with depth have been generated from the same amount of poses for the learning set, and 10,000 for the test set. In the generated depth images, the distance from the human model to the camera varies from 1.5 to 5.74 meters. Note that we used the specifications of the *Kinect v2* of *Microsoft* to generate the depth images and we added a Gaussian noise with the characteristics given in [4]. Some examples of our input depth images are shown in Figure 3 with the projection of the ground truth body joints positions in green.

We report the results obtained with 1, 4, and 12 models specialized according to the orientation. We analyze 8 body joints: neck, head, shoulder, elbow, wrist, hip, knee, ankle. We only consider the right joints given that the prediction accuracy will be symmetrical for the left ones.

Table 1. Mean errors on the positions of the considered body joints for different number of models used with a constant learning dataset size. There is an optimal number of models (4 in this experiment) for a constant learning dataset size (8000 samples).

	amount of models:	1		4		12
	learning samples per model:	8000		${}^{8000}/_{4} = 2000$		${}^{8000}/_{12} \simeq 666$
	range of each model:	360°		${}^{360^\circ}/_{4} + 2 \times 10^\circ = 110^\circ$		${}^{360^\circ}/_{12} + 2 \times 10^\circ = 50^\circ$
mean error	neck	2.9 cm	>	2.4 cm (- 15.3 %)	<	2.4 cm (- 14.4 %)
	head	3.1 cm	>	2.8 cm (- 7.6 %)	<	2.9 cm (- 3.9 %)
	right shoulder	5.4 cm	>	3.0 cm (- 45.1 %)	<	3.0 cm (- 44.1 %)
	right elbow	9.1 cm	>	5.9 cm (- 35.3 %)	<	6.0 cm (- 34.0 %)
	right wrist	13.7 cm	>	9.9 cm (- 27.3 %)	<	10.3 cm (- 24.4 %)
	right hip	4.2 cm	>	2.8 cm (- 34.0 %)	<	2.8 cm (- 33.6 %)
	right knee	5.8 cm	>	4.5 cm (- 23.4 %)	<	4.6 cm (- 21.8 %)
	right ankle	8.3 cm	>	6.2 cm (- 25.5 %)	<	6.3 cm (- 23.9 %)

3.1 Improvement with a constant global learning dataset size

Our first experiment shows what happens when we increase the number of models, with smaller orientation ranges, while keeping a constant learning dataset size. Table 1 gives the mean Euclidean errors for 1, 4 and 12 models. We see a significant reduction of the error for all joints when going from 1 to 4 models. These results underline that using multiple models designed for narrow ranges of orientations is preferable than using a unique model. However, going from 4 to 12 models slightly worsens the performance.

With a learning dataset, whose size cannot be increased, there is a trade-off between, on the one side, the improvement that is obtained from the knowledge of an approximative orientation estimation by the use of specialized pose estimation models, and on the other side, the deterioration due to the reduction of the learning set size. Nevertheless, the optimal solution takes advantage of a few models, and benefits from the knowledge of the orientation.

Note that the predictions for the head and the neck are less influenced by the number of models used. Indeed, the joints on the spine (that is the person's rotation axis) are less affected than those in the limbs by a change of the orientation. Moreover, we observe the largest errors on the wrist, as it is the joint that has the higher freedom to move in space. The magnitude of the mean error is thus related to the variety of poses in the test set. The general trend is higher errors at limb extremities, and lower errors at joints close to the torso.

The curves of Figure 4 depict the mean Euclidean errors (estimated with a Gaussian filter of $\sigma = 8°$) affecting the pose estimation at every joint with respect to the orientation of the observed person. The results obtained with a single 360°-model is shown in red, while the one with four 110°-models is shown in blue. As can be seen, the errors are anisotropic, and the best improvement obtained thanks to our principle is for people facing the camera, or seen from their back. Moreover, for all the joints of the right limbs, we observe larger errors when the person is seen from his left side, which is probably due to the fact that these joints have a higher chance of being occluded.

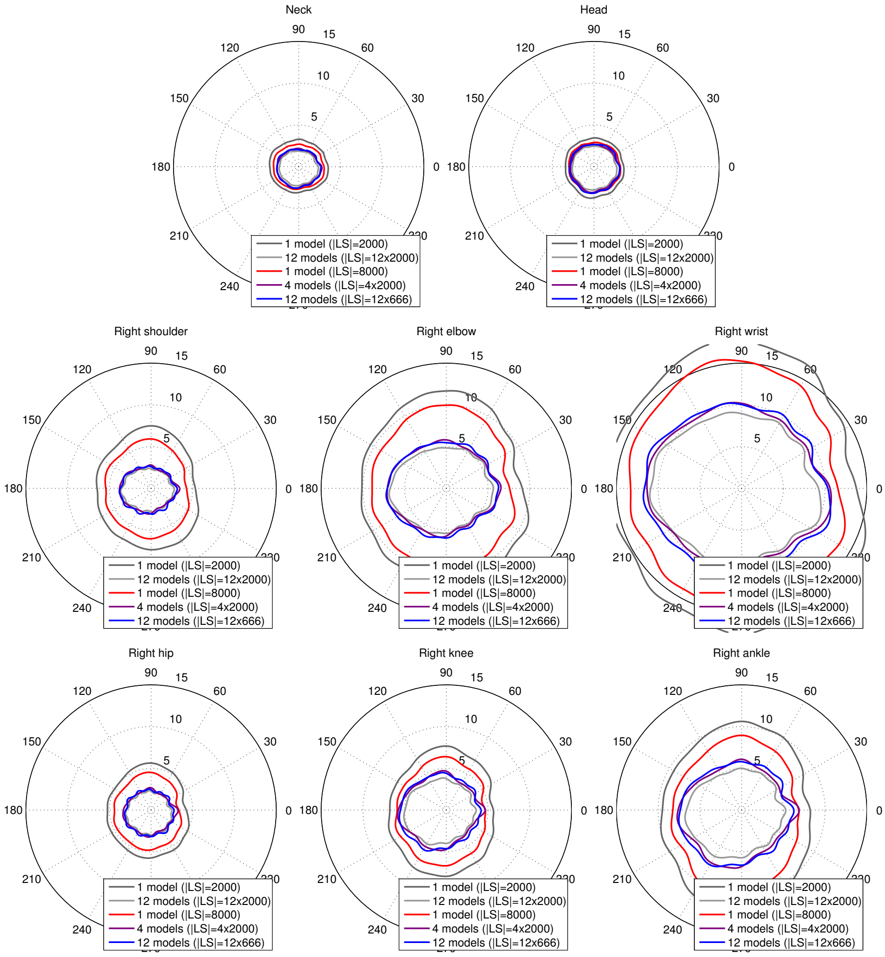

Fig. 4. Mean errors (in cm) on the positions of the considered body joints for different number of models specialized for reduced ranges of orientation. By convention, a person with an orientation of 0° is seen from the right side.

3.2 Improvement with a constant learning dataset size per model

Figure 4 also shows the behavior when the same experiment is performed with all models derived from the same amount of learning samples. The dark gray curves correspond to a single 360°-model, the purple ones to four 110°-models, and the light gray ones to twelve 50°-models. Each of these models has been learned from 2,000 samples. To the contrary of our first experiment, we observe a systematic decrease of the error when the number of models is increased. However, the small difference between 4 and 12 models suggests a plateau is reached after 4 models.

Therefore, relying on too many models is useless. This suggests that a rough orientation estimation suffices to improve the performance of pose estimation.

4 Conclusion

This work presents the principle of using an estimation of the orientation of the observed person to improve the accuracy of a pose estimation algorithm. Instead of learning a unique model over the 360°-range of orientation, we learn several models designed for smaller ranges of orientations. We tested this principle for different amounts of models and showed that the accuracy is significantly improved when the number of models increases while keeping a constant learning dataset size.

Acknowledgements. Samir Azrour is supported by a research fellowship of the Belgian National Fund for Scientific Research (F.R.S.-FNRS). Sébastien Piérard works on a project funded by the Walloon region of Belgium.

References

1. Carnegie Mellon University. Motion capture database. http://mocap.cs.cmu.edu.
2. P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learning*, 63(1):3–42, Apr. 2006.
3. R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Int. Conf. Comput. Vision (ICCV)*, pages 415–422, Barcelona, Spain, Nov. 2011.
4. C. Kerl, M. Souiai, J. Sturm, and D. Cremers. Towards illumination-invariant 3D reconstruction using ToF RGB-D cameras. In *Int. Conf. on 3D Vision (3DV)*, volume 1, pages 39–46, Tokyo, Japan, Dec. 2014.
5. S. Piérard, D. Leroy, J.-F. Hansen, and M. Van Droogenbroeck. Estimation of human orientation in images captured with a range camera. In *Advanced Concepts for Intelligent Vision Syst. (ACIVS)*, volume 6915 of *Lecture Notes Comp. Sci.*, pages 519–530. Springer, 2011.
6. S. Piérard, V. Pierlot, O. Barnich, M. Van Droogenbroeck, and J. Verly. A platform for the fast interpretation of movements and localization of users in 3D applications driven by a range camera. In *3DTV Conference*, Tampere, Finland, June 2010.
7. S. Piérard and M. Van Droogenbroeck. On the human pose recovery based on a single view. In *Int. Conf. Pattern Recogn. Applicat. and Methods (ICPRAM)*, volume 2, pages 310–315, Vilamoura, Portugal, Feb. 2012.
8. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Int. Conf. Comput. Vision and Pattern Recogn. (CVPR)*, pages 1297–1304, Providence, Rhode Island, USA, June 2011.
9. J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2821–2840, Dec. 2013.

10. X. Wei, P. Zhang, and J. Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM Trans. on Graph.*, 31(6):188.1–188.12, Nov. 2012.
11. K.-Y. Yeung, T.-H. Kwok, and C. Wang. Improved skeleton tracking by duplex Kinects: A practical approach for real-time applications. *Journal of Computing and Information Science in Engineering*, 13(4), Oct. 2013.