Monocular Estimation of 3D Poses from a Distance

Márton Véges, Viktor Varga

Abstract: Most 3D pose estimators only estimate egocentric coordinates where the body is centred at the origin. This is suitable for scenes with a single person but for images with interacting persons it is insufficient. We propose a monocular depth estimator for telephoto lenses to estimate 3D coordinates centred at the camera.

Our method fuses a depth map predictor and a relative 3D pose estimator by means of a 3-layer neural network. We compare the algorithm with the state-of-the-art method and show a 19% improvement.

Introduction

In activity recognition one of the important input features is the human skeleton positions in 3D or 2D space [3][12]. There are many methods to predict 3D positions from a single input image [6][2]. These methods usually only predict the relative coordinates of the body joints in an egocentric coordinate system. In remote situations involving multiple people, the global coordinates of the individuals are needed for estimation.

To this end, we use a monocular depth estimator that infers the distance of each pixel from the camera from a single image. We note that while this is an inherently ambiguous problem, there are multiple features in a scene that can help to infer the real coordinates of objects. Examples are tables, shadows or people whose size vary on a relatively small scale.

The task is difficult since (i) we are lacking depth information, (ii) there is a rich variety of body configurations, (iii) self-occlusion is frequent in monocular viewing, and (iv) small errors around the boundaries of the limbs can cause a joint placed in the background. These limitations can result in large prediction errors.

To mitigate the above problems we combine a depth estimator with a relative 3D pose estimator using a fusion network. We test the algorithm on the NTURGB-D database containing multiple people and find that the mean absolute error is below the half of the average distance between two persons over a variety of actions. These actions include hugging, object passing and handshakes among others.

We are aware of a single paper that targets the same problem [7]. Our method performs better on NTURGB-D by a large margin.

Related Work

Depth estimation Recent models use variations of convolutional neural networks. Laine et al. trains a fully convolutional model with residual up-projection blocks [5]. In [13] a fully convolutional network is trained using synthesised data from video sequences. The training process is unsupervised, all supervisory data is coming from synthetically generated frames.

3D pose estimation 3D pose estimators fall into two groups: methods in the first one predict 2D positions from images first and then predict the final 3D coordinates. Methods in the second one do the task in one step without estimating intermediate 2D positions. In the first category Martinez et al. [6] uses the Stacked Hourglass network [9] for 2D pose estimation and then applies a 5 layer fully connected network with residual connections to regress 3D coordinates. In the second category, Dabral et al. [2] train a network end-to-end applying structural losses on bone lengths and angles.

Global pose estimation Estimating global coordinates and not relative ones is a much less common objective in the literature. In [7] the authors apply a least squares minimisation between the predicted 3D and 2D coordinates. Vnect extends this to videos with additional loss components ensuring temporal smoothness [8].

Method

The proposed pipeline has four stages. First, we run the state-of-the-art OpenPose joint detector [1]. The result of the algorithm is the locations in pixels of 18 joints (in total $18 \cdot 2 = 36$ values) for each of the bodies in the image.

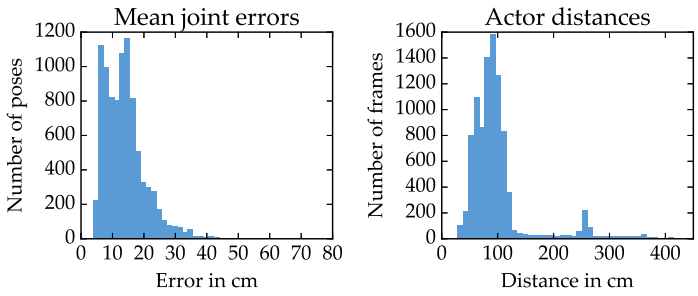

Figure 1: Left figure: distribution of the mean per joint error [4] of our model over all body poses. Right figure: distribution of the distances between the hips of actors on frames with two persons. All distances are in centimetres.

Second, we take the estimated 2D image coordinates of the joints and use the method of Martinez et al. [6] to convert them to relative 3D coordinates. At this stage, the hip is centred at the origin.

Third, a depth predictor trained on the NYU Depth V2 dataset [11] is run (referred as FCRN-D later on) [5]. It returns an estimation of the depth for each pixel in the input image. Note that the scale is ambiguous due to the difference between the calibration of the cameras that took the input image and the images in the training set.

Finally, we employ a fusion network to predict the final depth of the hip. The network has three dense layers with a residual block and batch normalisation applied for additional regularisation. The input is the image coordinates of the 18 joints generated in the first step and the corresponding depth values from the predicted depth map. The output is the distance of the hip from the camera. We get the final coordinates by calculating the coordinates of the hip in a camera centred coordinate system and moving the skeleton there.

Experiments

Database For the experiments the NTURGB-D database [10] was used. It contains 60 different actions performed by 8 actors taken in an indoor setting from 3 different camera angles. Since we are interested in multi-person interactions, we selected only the video sequences containing two persons. The data was split in training and test set by actors: videos of actors #1 and #2 formed the test data. Only camera setup #1 was used. In total the training set contained 370 videos, 25502 frames and 50404 body poses while the test contained 124 videos, 9890 frames and 19345 body poses.

The annotations of the database are based on Kinect that results in a couple of incorrect labels. To check whether these have any effect on the performance, we also ran the algorithm on a subset of the full database: only those body poses are kept where the ground truth joints were at most 40 pixels away from OpenPose's detection. 40 pixels corresponds roughly to 10 centimetres in real coordinates. Since OpenPose's accuracy is very high, it is a strong indicator whether an annotation was correct or not. This subset of the database is referred as Filtered DB later.

Error metric To evaluate the results we used the standard mean per joint position error which is the average Euclidean error over all joints and all poses [4].

Algorithms We compare our method to that of Mehta et al [7]. They minimise the projection error between the predicted 2D coordinates and the translated relative 3D coordinates. We also include a variation of our pipeline to show that the final stage fusion network is indeed needed. In this variation, the depth of the hip is estimated directly from the FCRN-D depth map. To overcome the issue of different calibrations, a simple linear regression is used between the predicted depths and the real distances from camera.

Table 1: Mean joint errors on the NTURGB+D database in centimetres. Best result selected in bold.

	Whole DB	Filtered DB
w/o Fusion	43.0	-
Baseline [7]	17.1	16.4
Ours	**13.8**	**12.9**

Discussion

The results of our experiments are summarised in Table 1. Our method improves 3.3 centimetres over the baseline which is a 19% decrease in the error, reaching state-of-the-art results. Since the average error hides a lot of details, we also present the histogram of the errors over all body poses in Figure 1 (left graph). The right graph in Figure 1 shows the distribution of the distances between the hips of the two actors on the images. The 5th percentile of interpersonal distances is 50.2cm while the median is 88.4cm. The median and 95th percentile of average errors are 13.1cm and 26cm respectively. Hence, we can tell the difference between the people in the images in around 95% of the cases.

We have also run ablation studies to analyse the components of our system. First, when using only the depth estimation of FCRN-D without our fusion network, we get an error of 43cm. This shows that a significant improvement is coming from the network and it is not only FCRN-D producing the results. Note, while the error of 43cm is far below the baseline it is still less than half of the median interpersonal distance.

Second, since the database contains erroneous annotations, we tested whether the improvements are only due to gains on the outliers or not. For this we used a subset of the full test set keeping only frames where the annotations were close to OpenPose's detections in 2D. The results show that our model is still better than the baseline (12.9 vs 16.4cm) indicating that learning mislabellings are not the root cause of the better performance.

Conclusions and Future Work

We have showed that our pipeline produces state-of-the-art global pose estimation on the NTURGB-D database. The lower error is not a result of the off-the-shelf components or incorrect annotations.

Our method can be extended in several ways, e.g. handling outdoor scenes or occlusion of body parts by objects. Also, end-to-end learning is a popular approach where the components are not treated separately but trained simultaneously. It often produces superior results.

Acknowledgements

The authors would like to thank their supervisor, András Lőrincz, and also Áron Fóthi for their guidance and helpful comments. The project was supported by the European Union and co-financed by the European Social Fund (EFOP-3.6.3-16-2017-00001).

References

[1] Z. Cao, T. Simon, S-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1302–1310, 2017.

[2] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, and A. Jain. Structure-aware and temporally coherent 3d human pose estimation. arXiv:1711.09250, 2017.

[3] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2186–2200, Nov 2017.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[5] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *The Fourth International Conference on 3D Vision*, pages 239–248, 2016.

[6] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *The IEEE International Conference on Computer Vision*, pages 2659–2668, 2017.

[7] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *The Fifth International Conference on 3D Vision*, 2017.

[8] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):44, 2017.

[9] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

[10] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760, 2012.

[12] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):914–927, May 2014.

[13] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 6612–6619, 2017.