Exploiting temporal context in 2d to 3d human pose regression

Viktor Varga, Márton Véges

Abstract: A major drawback of end-to-end image to 3d pose estimation approaches is the abscence of rich, in-the-wild image datasets with 3d human pose annotation. In this paper we show, that splitting the task and solving the subproblems of image based 2d pose estimation and 2d-to-3d coordinate regression independently is a viable approach. What is more, we present a lightweight deep learning based model to perform 2d-to-3d human body pose regression that is able to exploit temporal information and thus improve the state of the art.

Introduction

Automated understanding of human interactions in public spaces is both challenging and important. It is imperative to have the ability to reconstruct the 3-dimensional spatial model of the participants, to reliably estimate the details of such an interaction. However, usually, we only have access to monocular images or videos of the subject. In this case, human motion is perceived through a 2-dimensional projection: restoring the depth of points is an underdetermined problem. Ambiguity of the solution remains, even if we add different geometrical constraints derived from the anatomy of the human body.

In the past few years several end-to-end solutions were given to the problem of 3d pose estimation from monocular images [1]. Since image datasets with 3d pose labels are scarce and usually they were recorded in controlled environments, it becomes challenging to train end-to-end 3d pose estimators which generalize well. Another approach is to independently solve the subproblem of 2d pose estimation from images and then predict the 3d pose from the 2d coordinates. Image databases with 2d human pose annotation are abundant, including many in-the wild data, probably this is why the problem of 2d pose estimation from images has been well studied.

The ambiguity of restoring the 3d coordinates from a 2d projection may be reduced by using videos as input, or by adding temporal constraints based on the dynamics of the human body.

In our paper we present an image to 3d pose estimation pipeline, which exploits the temporal structures behind the data. To achieve this, we utilize a state-of-the-art image to 2d pose estimation software. Our contribution is the remaining part of the pipeline, namely a deep learning solution which estimates 3d pose coordinates from a series of 2d coordinates. Our method improves the state of the art by almost 15% in the former subtask.

Related Work

2d to 3d joints Lee and Chen were among the first to deal with 2d to 3d pose coordinate regression [2], interpreting the task as a binary decision problem for each limb. Several papers introduce various constraints based on the structure and dynamics of the human body: Dabral et al. [1] propose angle limits for valid limb configurations, Zhou et al. [3] move towards personalized pose estimation with their learnt limb-length ratios. Recently, Martinez et al. [4] show that lightweight, general deep learning models can outperform many complex solutions.

Exploiting temporal information Making use of the temporal context helps us to reduce noise and exploit the dynamics of the human body or the laws of physics. Zhou et al. [5] uses temporal smoothing on input 2d poses, Mehta et al. [6] penalizes velocity and acceleration.

Method

The task we aim to solve is 3d human pose estimation from 2d pose joint locations and their temporal context. Formally, we learn function $f^* = \min_f \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \|f(x_{t,j}) - y_{t,j}\|_2^2$, where $x_t \in \mathbb{R}^{2J}, y_t \in \mathbb{R}^{3J}$ are 2d and 3d body poses respectively, represented by a vector of joint points. J is the number of joints predicted and T is the length of the temporal window of analysis. The input and output of our model is a series of 2d and 3d body poses.

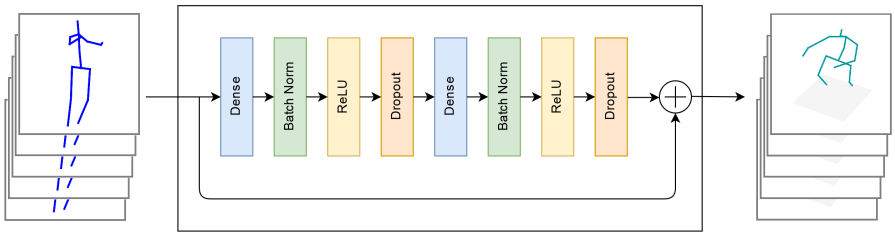

Figure 2: The architecture of our approach. The input is multiple vectors of 2d pose coordinates concatenated together. The concatenation of the corresponding 3d pose coordinates form the output of the model. The residual block at the center is repeated several times to achieve best results.

2-dimensional pose estimation We use the state-of-the-art 2d pose estimation software of Newell et al. [7] to make our pipeline capable of predicting pose in images, thus enabling the usage of the popular benchmark, the Human3.6M dataset [8] for us. We also use the 2d camera projections of the 3d pose coordinates to evaluate our 2d to 3d regression method independently.

DNN architectures Our approach exploits deep learning and its most recent results to achieve state-of-the-art prediction performance. We utilize deep neural networks with the well known ReLU nonlinearities [9]. We add batch normalization [10] and residual skip connections [11] to fight off the problem of vanishing gradients in very deep networks. Schematics of our architecture is shown in Fig. 2. This architecture was inspired by the works of Martinez et al. [4] and many others, who used similar combinations of the aforementioned layers and techniques for various purposes.

Loss function In our work, we focus on techniques that exploit temporal information and we evaluate the effect of three extra terms added to the standard L_2 loss function during the training procedure. We will refer to the first two as *Smoothness loss terms*, and define it as the mean square of the first and second order derivatives of the predictions over time. A somewhat similar term was introduced by Mehta et al. [6]. The third loss term is referred to as *Temporal limb length invariance term* which penalizes deviations of limb lengths over time in the predictions. This loss term is a novelty in its exact form, but can be derived from the work of Zhou et al. [3], who personalized predictions with fixed limb length ratios. The extra loss terms were added to the standard L_2 loss term with constant weights α, β, γ respectively.

Temporal smoothing Since our chosen 2d pose estimation software runs on single images, noise and outliers are expected to appear in its predictions. While our model may learn tasks as noise reduction or removal of outliers, we also evaluate our model on temporally smoothed input. We apply a low pass filter on the input data, to get rid of high frequency noise. Our choice is a Savitzky-Golay filter [12] fitting order 3 polynomials using a window length of 41.

Data preprocessing To prevent overfitting, we transform the 3d labels to the coordinate frame of all four cameras used in the Human3.6M dataset. Additionally, we normalize all input and output data by subtracting the mean and dividing by the standard deviation of the training data, per feature.

Results

Hyperparameter search Regarding the temporal length of the input and output, we have found that the ideal number of samples to concatenate is 8 pose vectors with a sampling frequency of 50Hz. The architecture that was found to perform best has the following parameters: a dense layer width of 4096 and consists of 4 residual blocks. Disabling residual skip connections caused serious deterioration of the results in all experiments.

3d pose regression Quantitative results are shown in Table 1. In the task of 2d to 3d pose regression using ground truth input data, our method improves the state of the art by 14.2% (6.5 mm). In case we use the 2d pose detections of the Stacked Hourglass [7] architecture, our results are beaten by Dabral et al.[1], but their model was also trained on Human3.6M image data, unlike ours, which gives their results a clear advantage.

The effect of the extra loss terms and smoothing The contribution of the loss terms were evaluated both independently and in unison. The *Smoothness loss terms* could improve 1.3 mm on average. The *Temporal limb length invariance term* improved 0.5mm on its own, and 1.5 mm together with the former terms. The following weights were found to be optimal: $\alpha = 2, \beta = 10, \gamma = 1.5$. Applying temporal smoothing on the input 2d pose data, our results were improved by a further 0.9 mm. The 95th percentile error results indicate that the amount of failed predictions were also reduced, greatly helped by input smoothing.

Table 1: Quantitative results on Human3.6M (in mm). Second column of the table shows the mean of the per-action mean joint error. Third column shows the mean of the per-joint 95th percentile errors. 2d GT: 2d projections of the 3d annotations were used as input data, SH: 2d pose input data was provided by the Stacked Hourglass network [7], 2d Tr: the 2d pose estimation model was trained or fine-tuned on the evaluation dataset unlike in our case, *: monocular approaches

	Mean error	**Mean 95th percentile error**
Martinez et al. [4] (2d GT) *	45.5	81.2
Ours (2d GT)	**39.0**	**70.5**
Martinez et al. [4] (SH) *	67.5	113.1
Zhou et al. [3] (2d Tr)	64.9	-
Dabral et al. [1] (2d Tr)	**52.1**	-
Ours (SH)	63.8	105.4
Ours (SH, smoothing)	62.7	100.9

Discussion and future work

In this paper we showed that a general deep learning approach solves the problem of 2d to 3d human pose regression effectively and can exploit temporal structures hidden in the data. Still, the surprisingly good results of Dabral et al. [1] show that an end-to-end approach may be even stronger. While the decoupling of the image-to-2d-pose estimation subproblem enables us to select our datasets from a much broader and richer palette, many precious image features, that let us infer depth, are lost.

As we have seen in the precious section, the application of a low pass filter over the input data resulted in an improvement. When we removed those samples from the Stacked Hourglass 2d pose predictions which contained clear failures (values over 100 mm in the second derivative of the time series of any joint when using a sampling rate of 20 milliseconds), the 3d pose estimations of our method improved a further 3.5 mm. We attribute this to the sensitivity of the low pass filter to extreme outliers. Instead, the application of Robust Principal Component Analysis (RPCA) [13] seems more appropriate, since this technique is less prone to highly corrupted data [14]. The investigation of this alternative is another potential future research task.

Acknowledgements

The authors are grateful to András Lőrincz, their supervisor, and Áron Fóthi, fellow researcher for their guidance and help. The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00001).

References

[1] Dabral, R., et al. "Structure-Aware and Temporally Coherent 3D Human Pose Estimation." *arXiv preprint arXiv:1711.09250* (2017).

[2] Lee, H.-J., et al. "Determination of 3D human body postures from a single view." *Computer Vision, Graphics, and Image Processing*, 30.2 p. 148-168 (1985).

[3] Zhou, X, et al. "Weakly-supervised Transfer for 3D Human Pose Estimation in the Wild." *arXiv preprint arXiv:1704.02447* (2017).

[4] Martinez, J., et al. "A simple yet effective baseline for 3d human pose estimation." *IEEE International Conference on Computer Vision* Vol. 206, p. 2640-2649 (2017).

[5] Zhou, X., et al. "Sparseness meets deepness: 3D human pose estimation from monocular video." *IEEE conference on Computer Vision and Pattern Recognition.*, p. 4966-4975 (2016).

[6] Mehta, D., et al. "Vnect: Real-time 3d human pose estimation with a single rgb camera." *ACM Transactions on Graphics (TOG)*, 36.4: 44 (2017)

[7] Newell, A., et al. "Stacked hourglass networks for human pose estimation." *European Conference on Computer Vision*, p. 483-499 (2016).

[8] Ionescu, C., et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." *IEEE TPAMI*, p. 1325-1339 (2014).

[9] Nair, V., and Hinton, G. E.. "Rectified linear units improve restricted boltzmann machines." *Proceedings of the 27th International Conference on Machine Learning*, p. 807-814 (2010).

[10] Ioffe, S., et al. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *International Conference on Machine Learning*, p. 448-456 (2015).

[11] He, K., et al. "Deep residual learning for image recognition." *IEEE conference on Computer Vision and Pattern Recognition*, p. 770-778 (2016).

[12] Savitzky, A., and Golay. M. JE. "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry 36.8*, p. 1627-1639 (1964).

[13] Wright, J., et al. "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization." *Conference on Neural Information Processing Systems*, p. 2080-2088 (2009).

[14] Milacski, Z. A., et al. "Robust detection of anomalies via sparse methods." *International Conference on Neural Information Processing. Springer, Cham*, p. 419-426 (2015).