Dense Pose Tracking

Yuxiang Zhou, Jiankang Deng and Stefanos Zafeiriou
Imperial College London, London, UK

***Abstract*— We explore the use of shape-based representations as an auxiliary source of supervision for pose estimation. We show that shape-based representations can act as a source of 'privileged information' that complements and extends the pure landmark-level annotations. In this work, we use 2D shape-based supervision signals, such as Support Vector Shape and train a 2-stacked hourglass model in cascaded manner for human pose estimation.**

I. Method

A. Dense Body Pose Estimation Networks

Our work in dense body pose estimation networks are inspired by learning with 'Privileged Information' [10], [1], [2], where it is argued that one can simplify training through the use of an 'Intelligent Teacher' that in a way explains the supervision signal, rather than simply penalizing misclassifications. This technique was recently used in deep learning for the task of image classification [2]. It shows that shape-based representations provide an excellent source of privileged information for human pose estimation. This additional information is only available during training, only serves as a means of simplifying the training problem, and only requires landmark-level annotations, as all current methods do. Another way of stating this is that we use shape-based representations to construct a set of auxiliary tasks that accelerate and improve the training of pose estimation networks. Additional dense supervision signals used in training the model are Support Vector Shape (SVS) [9].

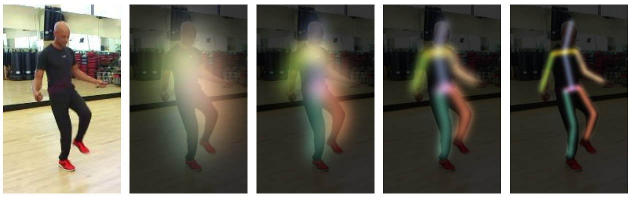

Fig. 1. Multichannel Support Vector Shape representations using different granularities. From left to right we show the SVS for $C = [3, 6, 12, 24]$ respectively, where C is the scaling of the underlying SVM data term.

1) Support Vector Shapes (SVS): A Support Vector Shape (SVS) is a decision function trained on binary shapes using Support Vector Machines (SVMs) with Radial Basis Function (RBF) kernels [7] - a shape is represented in terms of the classifier's response on the plane. This representation can be applied to both sparse landmark points and curves, fuses inconsistent landmarks into consistent and directly comparable decision functions, and is robust against noise, missing data, self-occlusions and outliers.

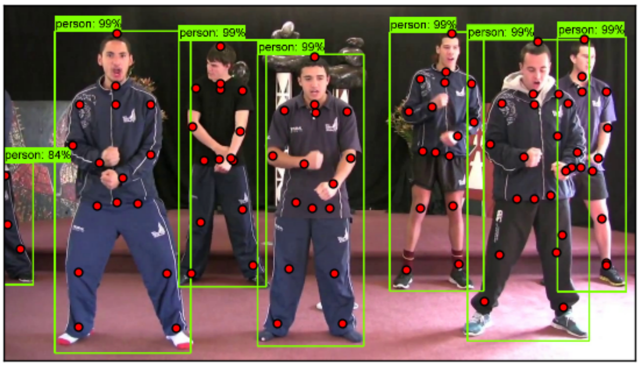

Fig. 2. Examplar predictions of estimated multi-person body joints correspondence with detection bounding boxes. Figure best viewed by zooming in.

The annotations for all training images are densely sampled to yield a set of landmarks per image. SVM training proceeds by assigning landmarks to the 'positive' class and randomly sampled points around them are assigned as belonging to the 'negative' class. SVMs with RBF kernel functions can map any number of data points onto an infinite-dimensional space where positive and negative points are linearly separable, hence the classification boundary on the 2D space represents the actual shape of the object. As in [7] we use class-specific losses to accommodate the positive/negative class imbalance problem. We extend the SVS representation to support also the case where multiple parts are annotated. It can provide further guidance on the estimation of dense shape correspondences for various object classes. In the case of human poses, 7-channel SVS are defined as in Figure 1.

B. Network Architecture

This section provides some details regarding our network architecture used to perform prediction of body poses and landmarks. In particular, we built our architecture based on the stacked hourglass networks, which is originally proposed in [6]. It consists of a series of convolutions and downsampling, followed by a series of convolutions and upsampling, interleaved with skip connections that add back features from high resolutions. The symmetric shape of the network resembles a hourglass, hence the name.

In our architecture, 2 hourglasses are stacked. The first hourglass is used to regress to dense shape information while the second one takes as input the image and the privileged information and regresses to landmark locations.

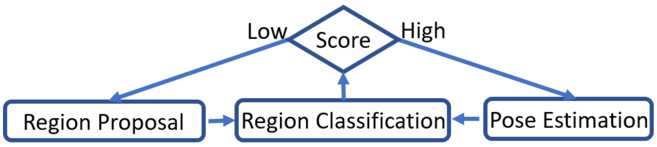

Fig. 3. Human detection, pose estimation and tracking framework.

Regarding loss function, we used pixel-wise ℓ_1 smooth loss for regressing to SVS signals that has continuous values.

C. Detection And Tracking

We train a Faster R-CNN [8] based human detector [4] on the Microsoft COCO dataset [5] with the 101-layer ResNet [3]. As shown in Figure 3, we illustrate the human detection, pose estimation and tracking framework. The Faster R-CNN detector consists of region proposal network and region classification network. After detection, human pose is estimated and the human region is updated by the predicted landmarks. The region classification network acts as the failure checker to interrupt the tracking. If the new human region has a high score, the pose estimation on next frame is initialized by the updated human region. Otherwise, human region proposal network is called. In this way, we generate the track-let for each human from the video.

II. Results

The results are submitted on PoseTrack for both per-frame multi-person pose estimation I and multiple object tracking II.

	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
our method	61.8	58.4	45.4	35.7	48.6	40.1	32.2	47.1

TABLE I
Per-frame multi-person pose estimation performance (AP).

	MOTP Total	Prec Total	Rec Total
our method	21.0	23.1	73.7

TABLE II
Multiple Object Tracking (MOT) metrics.

References

[1] Unifying distillation and privileged information. In *ICLR*, 2016.
[2] Y. Chen, X. Jin, J. Feng, and S. Yan. Training group orthogonal neural networks with privileged information. *CoRR*, abs/1701.06772, 2017.
[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
[4] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv preprint arXiv:1611.10012*, 2016.
[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
[6] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.
[7] H. V. Nguyen and F. Porikli. Support vector shape: A classifier-based shape representation. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.
[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
[9] H. Van Nguyen and F. Porikli. Support vector shape: A classifier-based shape representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):970–982, 2013.
[10] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 2009.