Improve Accurate Pose Alignment and Action Localization by Dense Pose Estimation

Yuxiang Zhou[1], Jiankang Deng[1] and Stefanos Zafeiriou[1,2]
[1] Department of Computing, Imperial College London, United Kingdom
[2] Centre for Machine Vision and Signal Analysis, University of Oulu, Finland
{*yuxiang.zhou10, j.deng16, s.zafeiriou*}*@imperial.ac.uk*

***Abstract*— In this work 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. We explore 2D shape-based supervision signals, such as Support Vector Shape. Our experiments show that shape-based supervision signals substantially improve pose alignment accuracy in the form of a cascade architecture. We outperform state-of-the-art methods on the MPII and LSP datasets, while using substantially shallower networks. For action localization in untrimmed videos, our method introduces additional classification signals based on the structured segment networks (SSN) and further improved the performance. To be specific, dense human pose and landmarks localization signals are involved in detection progress. We applied out network to all frames of videos alongside with output from SSN to further improve detection accuracy, especially for pose related and sparsely annotated videos. The method in general achieves state-of-the-art performance on Activity Detection Task for ActivityNet Challenge2017 test set and witnesses remarkable improvement on pose related and sparsely annotated categories e.g. sports.**

I. Introduction

Activity Detection and temporal action localization [25], [22], [13], [15], [7], [4], [14] has drawn increasing attention to the research community in past few years. Human activity understanding in untrimmed and long videos, especially, are crucial part of real-word applications including video recommendation, video surveillance, human-machine interaction and many others. It is of importance for algorithms to determining not only actions contained in videos but also temporal boundaries (activity starting/ending frames).

However, many methods are trained on short video clips where actions are tightly cropped, while, in practical, videos tend to be long and untrimmed. Also, spatial attentions for human poses are often neglected. The current methods of choice for human pose estimation are Deep Convolutional Neural Network (DCNNs). The general architecture applied for the task involves finding the parameters of a DCNN, which maps the image pixels to the locations of the body-parts. Recently, in order to incorporate contextual information in a better manner, cascade structures are proposed. In these cascade structures instead of regressing directly to part locations, fully convolutional networks are trained that output part detection heatmaps by representing each part using a circle/disc of a particular radius. Then, regression is performed on these heatmaps. The circle/disc representation could be sub-optimal for describing the parts, as well as the human body shape.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

Fig. 1. Frames in target video are extracted and applied human detector to generate human proposals before applying dense pose estimation to generate heat maps of landmarks and densely correspondent features. These additional signals are combined with the pyramid of action frame proposals from SSN to further improve the performance of determine starting/ending frames. Regarding timelines, **TOP** (blue) shows grondtruth action frame; **BOT** (green) shows predicted action frame. Figure best viewed by zooming in.

Driven by ActivityNet [4], a large-scale video benchmark for human activity understanding is released to the research community and consist of 200 activity categories, in which each contains 100 videos collected "in-the-wild". This dataset brought notable challenges to existing state-of-the-art approaches. In this paper, we integrated spacial attentions from human detection and dense human poses to further improve accuracy of both pose alignment and action localization. The contributions can be summarize as following:

- We propose a better way to use the contextual, as well as the body shape information by learning heatmaps that correspond to the support vector description of human body. We demonstrate that by using the proposed representation we train DCNNs that not only achieve state-of-the-art performance for body-part detection but also can recover the shape of the body as a whole.
- We migrated additional classification signals from the dense human poses above with spacial attentions alongside with the usage of structured segment networks (SSN) [25] as shown in Figure 1. Specifically, human body detection, semantic human body segmentation and pose landmarks localization signals are involved

in activity detection progress. We apply the additional signals to all frames of videos to generate dense human pose features and combined with results from SSN to further improve accuracy, especially for pose related and sparsely annotated videos as described in Fig 4. The method in general achieves state-of-the-art performance on test set and observed remarkable improvement on pose related and sparsely annotated categories.

II. Dense Body Pose Estimation Networks

Our work in dense body pose estimation networks are inspired by learning with 'Privileged Information' [19], [1], [5], [26], 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 [5]. 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 action detection task are Support Vector Shape (SVS) [18].

A. 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 [10] - 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.

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 [10] 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 2.

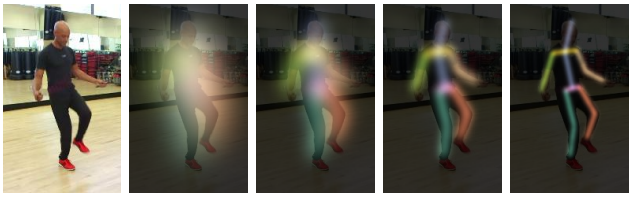


Fig. 2. Multichannel Support Vector Shape representations using different granularity. From left to right we show the SVS for $C = [3, 6, 12, 24]$ respectively, where C is the scaling of the underlying SVM data term.

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 [9]. It consists of a series of convolutions and down sampling, followed by a series of convolutions and up sampling, interleaved with skip connections that add back features from high resolutions. The symmetric shape of the network resembles a hourglass, hence the name.

In the stacked hourglass paper the best body pose estimation results have been achieved by using 8 hourglass networks. In our architecture, instead of 8, only 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. Regarding loss function, we used pixel-wise ℓ_1 smooth loss for regressing to SVS signals that has continuous values.

III. Activity Detection

The task for activity detection is to localize the temporal boundary of an activity. There are two types of action annotations as shown in Fig 4, a) action duration is long and almost one action per video in the category e.g. Zumba; b) action duration is short and one video contains multiple actions e.g. Long Jump. Most algorithms tackles the former situation well as these actions are covering majority part of the videos and their activity boundaries are very close to the starting/ending of videos. However, it is challenging to get sufficient accuracy for the second situation where most modern methods failed to provide accurate proposal.

Our methods are focused on determine accurate short activity boundary by incorporating attentions from human detector, poses landmark localization and pose segmentation with activity and completeness classifiers from SSN.

A. Structured Segment Networks (SSN)

The SSN network [25] relies on a proposal method (described in section III) to produce a set of temporal proposals of varying duration, where each proposal comes with a starting and an ending time. Given an input video, a temporal pyramid will be constructed upon each temporal proposals. One proposal is divided into three stages namely *starting*, *course*, and *ending*. In additional to the *course* stage, another level of pyramid with two sub-parts is constructed. To form the global region representations, features from DCNNs are pooled within these five parts and concatenated together. The

Fig. 3. TOP:Exemplar joints localization on MPII and LSP test set. BOT: Exemplar predictions of estimated dense whole body joints correspondence. Figure best viewed by zooming in.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Newell et al., ECCV'16 [9]	98.2	96.3	91.2	87.1	**90.1**	87.4	83.6	90.9	**62.9**
Bulat&Tzimiropoulos, ECCV'16 [3]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7	59.6
Wei et al., CVPR'16 [21]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5	61.4
Pishchulin et al., CVPR'16 [12]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4	56.5
Our Model	**98.4**	**96.4**	**92.0**	**87.9**	89.5	**88.4**	**85.1**	**91.4**	62.7

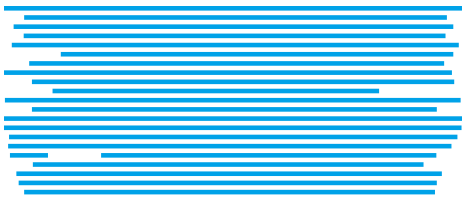
TABLE I

JOINTS LOCALISATION ACCURACY ON MPII DATASET.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Pishchulin et., ICCV2013 [11]	87.2	56.7	46.7	38.0	61.0	57.5	52.7	57.1	35.8
Wei et al., CVPR2016 [21]	**97.8**	92.5	87.0	83.9	91.5	90.8	89.9	90.5	65.4
Bulat&Tzimiropoulos, ECCV2016 [3]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7	63.4
Insafutdinov et al., ECCV16 [6]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1	66.1
Our Model	95.8	**95.8**	**94.9**	**92.3**	**95.2**	**96.6**	**95.7**	**95.2**	**69.6**

TABLE II

JOINTS LOCALISATION ACCURACY ON LSP DATASET.



(a) Zumba

(b) Long Jump

Fig. 4. Action annotations of two particular classes on the validation set. Each segment indicates the action duration, which is normalized by the whole video. Detection of the first action (Zumba) is much easier than the second action (Long Jump).

activity classifier and the completeness classifier operate on the the region representations to produce activity probability and class conditional completeness probability. The final probability of the proposal being positive instance is decided by the joint probability from these two classifiers.

B. Temporal Region Proposal

An input video is divided to 20 snippets and temporal region are generated based on sliding windows [14], [23]. A sliding window of size 3 are selected so 18 region proposals are generated. As we incorporated human detector [24], dense pose estimation and alignment as additional feature, existing proposed regions will be duplicated if the human pose detector returns multiple entries. For each proposed region, K-level temporal pyramid where each level dividing the region into smaller parts.

C. Temporal Region Classifiers

Structured temporal pyramid pooling [25] is performed to extract global features in which our detection, pose estimation and segmentation are involved. The training and testing of the classifiers are following the SSN network in similar manner.

Two types of linear classifiers (activity classifier and completeness classifier) are implemented on top of high-level features. Given a proposal, the activity classifier will produce a vector of normalized responses via a softmax layer which represents conditional distribution $P(c_i|p_i)$ where c_i is class label and p_i represents given proposal. The completeness

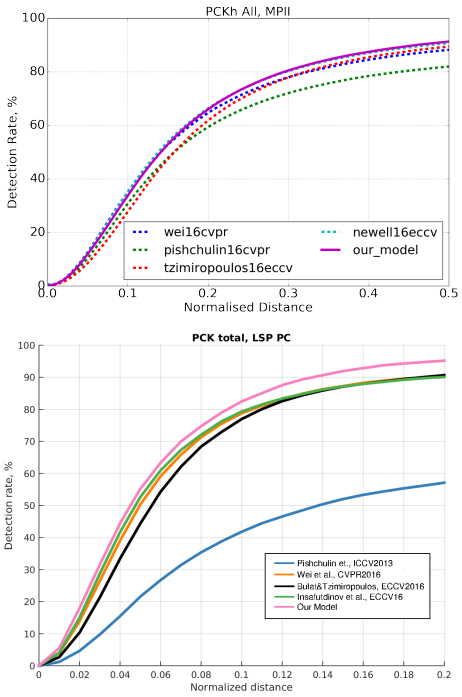


Fig. 5. PCKh plot of landmark predictions accuracy on **MPII (TOP)** test set and on **LSP (BOT)** test set. Out model using SVS signal is reported and compared to current state-of-the-art methods.

classifier C_k are trained for every activity class k. It can be expressed as $P(b_i|c_i, p_i)$ where b_i is binary indicator of the completeness of given region p_i. Outputs of both formed a joint distribution so the loss function is defined as:

$$Loss = -logP(c_i|p_i) - 1_{(c_i>1)}P(b_i|c_i, p_i)$$

where completeness term is active only when class label is not background.

IV. Experiments

A. Joints Localization on MPII & LSP

Databases The experiments outlined in this section are performed on two well known body pose databases: MPII Human Pose [2] and Leeds Sport Poses (LSP) + extended training set [8]. There are around 18k training images and 7k testing images involved in MPII. In section II-A, we split training set randomly to make a 3k size validation set while the rest are used for training. In section IV-B, the same 15k training set are used for training and reporting results of MPII. Results on LSP are reported by fine tuning the same model with the 11k extended LSP training set.

Evaluation Metrics The accuracy reported follow the Percentage Correct Keypoints (PCK) measurement on LSP dataset. Normalized PCK measurement by the scale of head (PCKh) is used for MPII on both validation and test set. Note that the performance gap between validation and test set is due to the use of invisible parts in measuring the performance. That is, in the validation set we measured the performance making use of the invisible parts, while the test set protocol of MPII does not use the invisible parts when computing the evaluation metrics.

Model Training Our model is implemented using TensorFlow [1]. 15k images from the training set mentioned above are used with augmentations. Each pose instance in the image was cropped to size 384×384. Cropped images are then randomly flipped, rotated by $\pm 30^{\circ}$ and rescaled by 0.75 to 1.25 before cropping to size 256×256. The model are trained with initial learning rate 1×10^{-3} with exponential decay factor of 0.97 at every 2 epochs. The models were trained for 100 epochs before testing.

Results reported on MPII are obtained by using SVS supervision signal. Figure 5 plot the cumulative error distribution on MPII test set and Table I summarize the quantitative results. Comparison with the state-of-the-art on LSP dataset is shown in Figure 5 and Table II.

Some qualitative results are collected in Fig 3 for the test sets of MPII and LSP. Top three rows show joint localization in challenging poses e.g. extreme viewing angles, challenging poses, occlusions, self occlusions and ambiguities. Bottom row demonstrates dense shape correspondence estimated on test images using SVS signal.

B. Action Localization

The results of our methods on ActivityNet 2017 (test set) are shown in Table III. Highest ranked results submitted to previous ActivityNet 2016 Challenge and current ActivityNet 2017 Challenge are involved in the table.

ActivityNet 2017 (Test Set)	
Method	Avg. mAP
Wang, R. and Tao, D. [20]	14.62
Singh, B. and Marks, T. et. al. [16]	16.68
Singh, G. and Cuzzolin, F. [17]	17.83
Zhao, Y. and Xiong, Y. et. al. [25]	28.28
Xiong, Y. et. al. [ActivityNet 2017]	31.863
Lin, T. et. al. [ActivityNet 2017]	33.406
Our Method	31.826

TABLE III
ACTION DETECTION RESULTS ON ACTIVITYNET 2017. EVALUATED BY MEAN AVERAGE PRECISION (MAP).

V. Conclusion

In this work, we have shown that shape-based representations can largely help CNN training for human pose estimation through the construction of auxiliary tasks that largely complement the landmark-level annotations. Our results indicate that we thereby accelerate training, improve accuracy, and outperform state-of-the-art deeper architectures on challenging benchmarks, while at the same time obtaining a network that not only provides landmarks but also delivers dense body joints. In future work we intend to further explore the use of shape-based representations for general object recognition.

VI. Acknowledgement

This work was partially funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 688520 (TeSLA) and by EPSRC Project EP/N007743/1 (FACER2VM).

[1] https://tensorflow.org

REFERENCES

[1] Unifying distillation and privileged information. In *ICLR*, 2016.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[3] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*. Springer, 2016.

[4] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[5] Y. Chen, X. Jin, J. Feng, and S. Yan. Training group orthogonal neural networks with privileged information. *CoRR*, abs/1701.06772, 2017.

[6] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision (ECCV)*, May 2016.

[7] Y. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014.

[8] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[9] 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.

[10] H. V. Nguyen and F. Porikli. Support vector shape: A classifier-based shape representation. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[11] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[12] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. *arXiv preprint arXiv:1703.01515*, 2017.

[14] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016.

[15] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[16] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2016.

[17] G. Singh and F. Cuzzolin. Untrimmed video classification for activity detection: submission to activitynet challenge. *arXiv preprint arXiv:1607.01979*, 2016.

[18] 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.

[19] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 2009.

[20] R. Wang and D. Tao. Uts at activitynet 2016. 2016.

[21] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[22] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. *arXiv preprint arXiv:1703.07814*, 2017.

[23] J. Yuan, B. Ni, X. Yang, and A. A. Kassim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2016.

[24] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, et al. Crafting gbd-net for object detection. *arXiv preprint arXiv:1610.02579*, 2016.

[25] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, D. Lin, and X. Tang. Temporal action detection with structured segment networks. *arXiv preprint arXiv:1704.06228*, 2017.

[26] Y. Zhou, E. Antonakos, J. A. i medina, A. Roussos, and S. Zafeiriou. Estimating correspondences of deformable objects "in-the-wild". In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16)*, Las Vegas, NV, USA, June 2016.