PERSONALIZED POSE ESTIMATION FOR BODY LANGUAGE UNDERSTANDING

Zhengyuan Yang, Jiebo Luo

Department of Computer Science
University of Rochester, NY 14627, USA
{zyang39, jluo}@cs.rochester.edu

ABSTRACT

To achieve high accuracy and stability in human pose estimation from videos, we propose a personalized model with a specially designed ConvNet structure and a visual similarity based iteration step. This model consists of: 1) a fully convolutional network with spatial fusion architecture to boost the accuracy of single-frame based joint predictions, 2) optical flow-based refinement to incorporate motion and temporal information, and 3) iterative personalized annotation to boost the reliability of the joint predictions. For benchmarking, our model outperforms the state-of-the-art on the public pose estimation datasets Chalearn and FLIC. Moreover, our model performs the best on a new psychiatric conversation dataset for computer vision based body language and emotion study.

Index Terms— Pose Estimation, Body Parts Tracking, Convolutional Neural Networks, Medical Video Analysis

1. INTRODUCTION

Human pose estimation is crucial for many applications. Traditional human pose estimation methods include edge-based histograms [1], silhouette contours [2], pictorial structures [3, 4, 5] and deformable part models [6]. The recent success of convolutional neural networks [7] in human pose estimation has greatly improved the overall performance of many systems [4, 8, 9, 10, 11, 12, 13, 14]. Significant improvements in accuracy have been observed on various datasets, including the FLIC dataset [15], the BBC Pose dataset [16, 17] and the Chalearn dataset [18]. However, with the additional challenge introduced by dramatic whole body movements (e.g. dancing, exercising) in these datasets, even the state-of-the-art methods can generate implausible predictions on certain joints [13]. Motivated by our targeted application of inferring subtle body languages and emotions from videos, we propose a personalized model to boost the accuracy and reliability of human pose estimation in videos with subtle human gestures.

In our task, a vision based model is desired to assist psychiatrists in assessing mental disorders from their conversation videos with patients. In conversation videos, subjects' subtle gestures pose challenges to the approach of direct action recognition with visual representations, and may cause frequent recognition failures. For example, touching nose and touching mouth with hands have similar visual perceptions, while they have totally different implications from the perspective of psychiatry. Some examples of major body languages can be found in Enkivillage [1]. As extract body langauges, pose estimation is the first step. The output results are then employed to infer subtle human body languages in psychiatric conversation videos.

To meet the requirements of high accuracy and reliability in human pose estimation, we propose the following improvements to the traditional ConvNet based methods. Inspired by [12], we propose a spatial fusion architecture to extract the spatial relations between the joints, while constructing the network as fully convolutional. Note that this is an alternative to learning spatial relations with graphical models [8]. Furthermore, we use an iterative personalized annotation step to boost the reliability of joint predictions. This step iteratively discards and regenerates joint predictions with a selection criterion based on the visual similarity of the joints throughout the entire video.

For performance evaluation, experiments are done on two public datasets, FLIC and Chalearn. The high accuracy verifies the effectiveness of the spatial fusion ConvNet architecture and the personalized annotation step. Furthermore, a psychiatric conversation dataset is established and tested on. Our model also outperforms the state-of-the-art on the proposed dataset and shows the promise for medical applications.

In summary, we make the following contributions:

- We propose a fully convolutional spatial fusion architecture, in contrast to graphical models, to encode the spatial relations between the joints.
- We introduce a personalized annotation step to boost the reliability of the joint predictions. The percentage of implausible annotations is iteratively reduced.
- We establish a psychiatric conversation dataset for computer vision based subtle body language extraction and mental disorder detection.

In the remainder of the paper, Section 2 describes our proposed model, Section 3 presents the datasets, the experimental design and the results, with conclusions in Section 4.

[1] http://www.enkivillage.com/body-language-examples.html

2. METHODOLOGY

2.1. Architecture Overview

In this paper, the psychiatric conversation video analysis task is formulated as an upper-body human pose estimation problem. Different from general pose estimation problems, this task has specific requirements on high accuracy and reliability in order to infer subtle body languages.

As shown in Fig. 1, our model includes three specifically designed components. A fully convolutional network with the spatial fusion layers takes individual RGB frames as the input and independently outputs predicted heatmaps for the frames. The output heatmaps are then refined with optical flow between the centered frame and its temporal neighbors. Finally, the visual features in patches around the joints are adopted to iteratively discard and regenerate joint predictions. We assume that joints and their surroundings should share similar visual representations throughout a long video.

2.2. ConvNet Model Architecture

We propose a fully convolutional architecture with spatial fusion layers. Instead of producing joint locations directly from RGB frames [9, 16], the network output is a 2D heatmap that represents the probability of the joints appearing at certain locations. The network architecture includes two parts, namely spatial layers and spatial fusion layers.

Spatial Layer Architecture The spatial layer architecture is constructed based on the model in [16], which contains five convolutional layers followed by two fully connected layers for joint coordinate regression. We replace the fully connected layers with three *additional convolutional layers*, converting the outputs from the directly regressed coordinates to the heatmaps of joint locations. The loss is calculated as the L2 distance between the predicted heatmap and the ground truth heatmap, which is generated as Gaussian distributions around the ground truth joint coordinates. This loss replaces the distance based loss used in the direct regression model. In a pose estimation task, the "likely correct predictions" might not be close to the ground truth. Instead, possible locations often locate near the patches that share the similar visual representations with the ground truth, for example the pair of the left and right wrists. The fully convolutional architecture can iteratively suppress the sub-peaks in the heatmaps, therefore help the network output better converge to the ground truth location. It is noteworthy that the heatmap output is convenient for visualization and further refinement.

Spatial Fusion Layer The fully convolutional spatial architecture allows the existence of sub-peaks in the heatmaps. Although this architecture helps the network better converge, it has a potential disadvantage in that the sub-peaks may be mistaken as the final prediction, for example, the prediction of the left wrist sometime appears near the right wrist. To solve this problem, five extra convolutional layers, namely spatial fusion layers, are proposed to learn the spatial relations among the joints. An alternative approach to learning spatial relations is using graphical models [8].

The input of the spatial fusion layers are the concatenation of Conv-Layer 3's output and Conv-Layer 7's output. The Conv-Layer 3's output contains more low-level visual features and information about the joint spatial relations, while Conv-Layer 7's output contains mostly the predicted confidence maps of the joint locations.

2.3. Optical Flow-based Refinement

Given the heatmaps predicted by the ConvNet, optical flow is then used to boost the accuracy and temporal smoothness of the predictions. For a target frame at time t, the frames in a time window $[t-n, t+n]$ are included for refinement. Optical flow is calculated between the centered frame and every other frames in the time window with DeepFlow [19]. The calculated optical flow is then warped with the heatmap predicted from the centered frame. Finally, the warped heatmaps are averaged with Gaussian weights and the joint locations are regressed as the locations of the maximum values in the warped heatmaps.

2.4. Personalized Annotation

To further boost the reliability of pose estimation and to avoid the occasionally implausible predictions, an iterative personalized annotation step is proposed. The idea of the personalized annotation step is to compare the similarity between the visual features around the joints in each frame and the ones in other frames. We make the assumption that the same patch's visual features should be similar in different frames in a long video. Therefore, existing joint annotations can be discarded and new candidates can be generated by matching the joint patches. The level of similarity is decided by a set of classifiers and the whole process is done iteratively, over the following four steps:

Initial Annotation: We generate the joint predictions with high reliability by setting a confidence value threshold in the step of converting the heatmaps to the joint coordinates. The network described in Section 2.2 and 2.3 is adopted.

Spatial Matching: We build a set of classifiers to match the joint patches from the annotated frames to ones in the candidate frames. Random forest classifiers are trained for individual joints with multiple window sizes for classification. For verification, HOG similarities are calculated between the candidate patches and the initially annotated patches.

Temporal Propagation: An exemplar-SVM is trained to match the patches with the visual similarity measurement. Based on the measurement and the optical flow calculated using the method in [19], new candidate annotations are propagated forward and backward to the neighboring frames. However, many are inaccurate and will be discarded during further iterations.

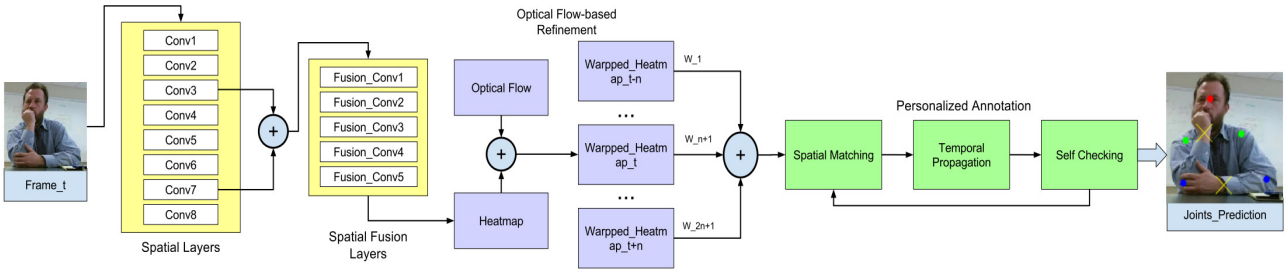

Fig. 1: Model architecture. Our proposed model contains three parts: a *fully convolutional network* for heatmap prediction, followed by *optical flow refinement*, and finally a *personalized annotation* step to boost the reliability of the predictions.

Self Evaluation: Annotation evaluation criteria are proposed to Automatically fuse and select the candidate annotations, including *a) Candidate Annotation Fusion*: If multiple predictions exist, which are propagated from different initial frames, a single prediction will be selected as the 2D location with maximum annotation density. Also, a pre-set annotation density threshold has to be met, or all the multiple predictions will be discarded; *b) Puppet Model*: We train a linear SVM with RGB and HOG features to determine if lower arms are in correct positions; *c) Occlusion Detection*: HOG and RGB features are fed into another SVM classifier for body parts occlusion detection. The annotations will be discarded if they fail to meet any of the evaluation criteria in the self evaluation step. In addition, failed joints will be randomly re-evaluated.

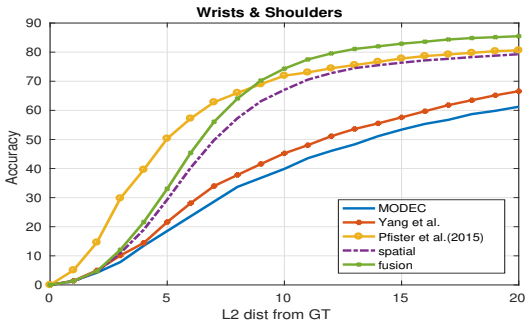

Fig. 2: Performance comparison on the FLIC dataset. Our experiments follow the standard approach on FLIC, reporting the accuracy of the wrists and shoulders with a normalized distance threshold (Torso height equals to 100 pixels).

3. EXPERIMENTS

3.1. Dataset

Frames Labeled In Cinema (FLIC) FLIC [15] is a dataset collected automatically from the Hollywood movies. The initial version of FLIC contains 5003 images, among which 1016 are selected as the testing set. Images are annotated manually with the crowd-sourcing marketplace Amazon Mechanical Turk. 10 upper-body joints are labeled and the median five labels are taken for each image.

Chalearn The Chalearn 2013 multi-model gesture dataset [18] includes the gesture data of 27 people, including audio, skeletal models, user masks, RGB and depth image sequences. There are 956 sequences, lasting 23 hours and include 1.3 million frames. Since the joint annotations are generated from Kinect, both training and testing labels are noisy. Chalearn shares several similarities with our target application because it is also collected by Kinect and focuses mainly on the upper body gestures.

Psychiatric Conversation Dataset This dataset is established in collaboration with the University of Rochester Medical Center (URMC). This dataset is built to assist studies on inferring subtle body languages from videos. Eight high-resolution long videos are collected, lasting from 15 to 30 minutes (9,000 to 18,000 frames). The content of each video is the conversation between a patient and a psychiatrist. Patients remain sitting behind a table throughout the video, and only the upper body is captured.

3.2. Component Evaluation

The evaluation method from [15] is applied, which calculates the accuracy based on a radius distance threshold. The input images are cropped based on the ground-truth torso height, and re-scaled to a height of 256. On the FLIC dataset, torso height is normalized to 100 pixels and on the psychiatric conversation dataset the normalized height is 150 pixels.

Spatial Fusion Layer On the psychiatric conversation dataset, we observe a significant improvement in accuracy with the spatial fusion architecture. There is an additional 3.5% gain from 74.2% to 77.7% at $d = 5$ pixels, and the gain remains in the high recall area from 90.0% to 93.5% at $d = 6$ pixels, where d is the radius distance threshold. Similar gains are also observed on FLIC and Chalearn. The improvement is obtained by learning the plausible joint locations with the extra spatial fusion layers, and suppressing the implausible predictions caused by visual ambiguity. Moreover, the deeper convolutional architecture further boosts the accuracy in the high precision area.

Optical Flow-based Refinement A further 1.6% gain is observed on the psychiatric conversation dataset from 77.7%

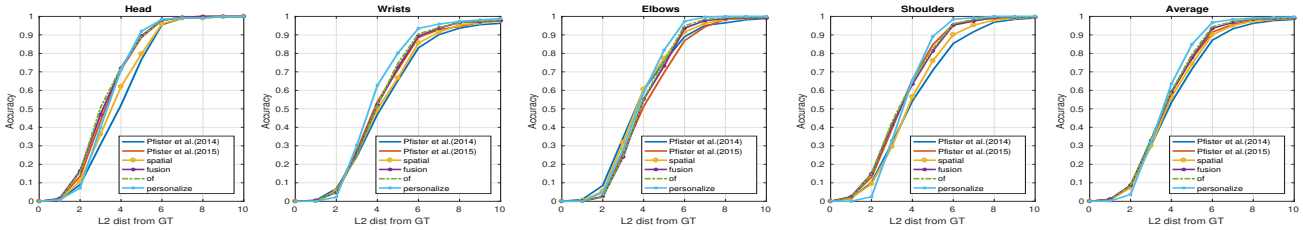

Fig. 3: Performance comparison on the psychiatric conversation dataset. Only the joints with annotations are used in accuracy calculation for the personalized annotation method. *Best viewed on screen with zoom-in, or in print.*

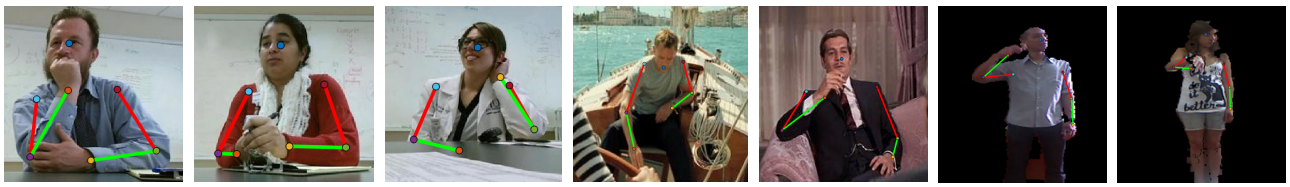

Fig. 4: Examples of pose estimation. The first three images and their predicted annotations are from the psychiatric conversation dataset. The middle two images are from FLIC. The last two images are from Chalearn. All the images shown above have been pre-processed (cropped, and background subtracted for Chalearn).

to 79.3% at $d = 5$ pixels. The optical flow-based refinement step considers the motion between frames and extract temporal information with a weighted average of the neighboring heatmaps. The refinement step improves the temporal smoothness of the predictions, and boosts the accuracy with information from the neighboring frames.

Personalized Annotation The iterative personalized annotation iteration step provides a significant gain of 5.6% on the psychiatric conversation dataset from 79.3% to 84.9% at $d = 5$ pixels. The iteration step greatly boosts the accuracy and reliability of the predictions, although it may fail to provide annotations to all the joints.

3.3. Comparison with the Baseline and State-of-the-Art

Psychiatric Conversation Dataset Our model is tested on the psychiatric conversation dataset. The results are presented in Table 1. Furthermore, Fig. 3 presents the accuracy curve under different distance thresholds. The model in [16] is compared as the baseline, which includes an end-to-end ConvNet to regress the location of the joints directly. Experiments are also conducted with the state-of-the-art method [12] on the psychiatric conversation dataset. The results are obtained using our re-implementation based on released partial code. We argue that our better performance is mainly due to the personalized annotation step. Although a small number of the predictions fail to be generated, the iterative personalized annotation step greatly improves the performance of the predictions, which is valuable for inferring subtle body languages.

FLIC/Chalearn Following the standard experiment approach on the FLIC dataset [15], we report the accuracy of the shoulders and wrists in Fig. 2. Size of the frames are normalized with the torso height to 100 pixels. The model in [20] are compared as the baseline. The results on Chalearn are similar and thus omitted to conserve space.

Method	Head	Wrsts	Elbws	Shldrs	Avg.
Pfister et al. [16]	76.8	65.4	75.6	70.9	71.5
Pfister et al. [12]	88.8	71.3	69.6	84.8	77.1
Spatial	79.8	66.8	77.4	76.0	74.2
Fusion	89.5	72.5	73.3	81.4	77.7
Fusion + OF	89.3	74.6	75.6	82.5	79.3
Fusion + OF + P	**92.0**	**80.2**	**81.8**	**89.2**	**84.9**

Table 1: Comparison of accuracy (%) on psychiatric conversation dataset with a fixed threshold at $d = 5$ pixels. The threshold is selected at which can best distinguish the difference between models.

4. CONCLUSIONS

We propose a novel model for highly accurate and reliable 2D human pose estimation. The model is a fully convolutional ConvNet with the spatial fusion layers. The output heatmaps are further refined with optical flow along with the visual representation around the joints. Our model outperforms several state-of-the-art methods on the proposed psychiatric conversation dataset. It also achieves the state-of-the-art accuracy on two public pose estimation datasets.

We plan to employ the predicted joints to infer body languages and emotion during psychiatric conversations, as well as combine pose estimation with action recognition.

Acknowledgment

We thank the generous support of New York State through the Goergen Institute for Data Science.

5. REFERENCES

[1] Greg Mori and Jitendra Malik, "Estimating human body configurations using shape context matching," in *European conference on computer vision*. Springer, 2002, pp. 666–680.

[2] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell, "Inferring 3d structure with a statistical image-based shape model," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 641–647.

[3] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3041–3048.

[4] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1347–1355.

[5] Marcin Eichner, Manuel Marin-Jimenez, Andrew Zisserman, and Vittorio Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International journal of computer vision*, vol. 99, no. 2, pp. 190–214, 2012.

[6] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[7] B Boser Le Cun, John S Denker, D Henderson, Richard E Howard, W Hubbard, and Lawrence D Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*. Citeseer, 1990.

[8] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in neural information processing systems*, 2014, pp. 1799–1807.

[9] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[10] Sijin Li, Zhi-Qiang Liu, and Antoni B Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 482–489.

[11] Sijin Li, Weichen Zhang, and Antoni B Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2848–2856.

[12] Tomas Pfister, James Charles, and Andrew Zisserman, "Flowing convnets for human pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1913–1921.

[13] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman, "Personalizing human video pose estimation," *arXiv preprint arXiv:1511.06676*, 2015.

[14] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[15] Benjamin Sapp and Ben Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *In Proc. CVPR*, 2013.

[16] Tomas Pfister, Karen Simonyan, James Charles, and Andrew Zisserman, "Deep convolutional neural networks for efficient pose estimation in gesture videos," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 538–552.

[17] J. Charles, T. Pfister, M. Everingham, and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos," *International Journal of Computer Vision*, 2013.

[18] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 445–452.

[19] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *IEEE Intenational Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013.

[20] Yi Yang and Deva Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1385–1392.