1+1>1? Quantitative Analysis of Multiple-Frame Effect for Human Pose Estimation

Jianfeng Xu[1,a)] Satoshi Komorita[1,b)]

Abstract: Although it is well known that the performance of recognition/classification can be effectively improved by integrating multiple sources (e.g., the accuracy of human pose estimation increases when using multiple frames rather than a single frame), it is very challenging to quantitatively analyze how much improvement will be obtained by an additional source and what factors will affect the performance improvement. As far as we know, this work presents for the first time a quantitative analysis of a particular case, where multiple frames are used to exploit temporal information for improving pose estimation in videos. More specifically, we select a cutting-edge technology, PoseWarper, as our analysis target. For simplicity, but without loss of generality, we focus on using two frames in PoseWarper. In this work, we not only discuss the necessary conditions for improving performance by using one more frame but also confirm that a linear regression works well to model the relationship between the accuracy gain and the time difference of two frames in the dataset of PoseTrack2017.

Keywords: temporal pose estimation, multiple-frame analysis

1. Introduction

In the tasks of recognition/classification, it is well known that the performance can be effectively improved by integrating multiple sources. For example, multimodal emotion recognition [21] fuses multiple relevant modalities including visual cues, audio cues and sensor data for better accuracy. Similarly, besides RGB images, depth data and IR images are helpful in the face recognition field [32]. Also, a two-stream approach [26] shows that an RGB stream and optical flow stream used together are effective in action recognition. In many cases, the performance has been demonstrated to be higher if more sources are integrated. Ensemble learning that includes boosted classifier like AdaBoost [12], [25] combines multiple "weak classifiers" into a single "strong classifier". The idea still works well even in the era of deep learning [22], when either XGBoost [6] or LightGBM [18] or CatBoost [24] is widely used. However, more memory and computational resource are consumed when more sources are used. With limited resources, it requires huge effort for experts to design or optimize the system so that the most effective source for the system is selected. Because there is no quantitative analysis available for this important issue in the literature, the system design/optimization is basically empirical requiring considerable know-how. Therefore, it is very important to know what will affect performance improvement and how much it can be improved when we use an additional source in terms of cost/performance balance.

In this work, we focus on a particular case, i.e., human pose estimation in videos [4], [17], [23]. As one of the successful applications for deep learning technologies, many powerful neural networks such as OpenPose [5], AlphaPose [11], CPN [7], and HRNet [28] were proposed for human pose estimation in still images, in which a heatmap is output for each joint of a person. Recently, new technologies have been further developed for videos, where an important issue is how to use temporal information in videos. As the state-of-the-art method, PoseWarper [4] is selected as our analysis target, which was the winner of PoseTrack2017 Challenge 2 "Multi-frame Person Pose Estimation" [1]. PoseWarper [4] estimates a warped heatmap from one frame to another. Therefore, multiple frames (e.g., five neighboring frames in the original paper [4]) can be fused into a heatmap, which achieves better performance than a single frame. There are many other papers [23], [27], [29], [30] that also use multiple frames to improve the accuracy of pose estimation in videos. These works lead to a natural question: how much accuracy gain we can obtain by using an additional frame.

For simplicity, but without loss of generality, we focus on using two frames in PoseWarper. Namely, we want to estimate the human poses in Frame t given Frame A and Frame B. Following the original paper of PoseWarper [4], suppose $\|t - A\| <= 2$ and $\|t - B\| <= 2$. In this work, we endeavor to answer the following questions.

- (qualitative analysis) Q1: Are there any conditions

[1] KDDI Research, Inc., Ohara 2–1–15, Fujimino, Saitma 356–8502, Japan
[a)] ji-xu@kddi-research.jp
[b)] sa-komorita@kddi-research.jp

whereby accuracy would be increased by using an additional frame? If so, what are the conditions?
- (quantitative analysis) Q2: What factors will affect the accuracy gain and in what kind of function?

For the first question, we observe that there is no guarantee that two frames will yield better performance than a single frame. Namely, it is possible that 1 frame + 1 frame <1 frame. Furthermore, in this work, we provide two necessary conditions for performance improvement (1+1>1). For the second question, by analyzing the experimental results of PoseWarper [4] in PoseTrack2017 [17], we demonstrate that a linear regression works well to model the relationship between the accuracy gain and the time difference of two frames. These are the main contributions of this work.

This work is organized as follows: in Section 2, we briefly introduce related work on human pose estimation in still images and videos. Then, in Section 3, we describe the necessary preparations for analysis including an introduction to the algorithm used in PoseWarper and the dataset of PoseTrack used in our experiments. Then, in Section 4, we report the details and our findings by analyzing the experimental results. Finally, in Section 5, we conclude this work.

2. Related Work

There has been significant interest in human pose estimation due to its importance in many applications such as pedestrian detection, understanding human behavior, sports analysis, and virtual reality. We give a brief overview of multi-person pose estimation in both still images and videos. For a detailed and complete survey, please refer to the survey paper [10].

2.1 Multi-Person Pose Estimation in Still Images

In the past decade, many papers on multi-person pose estimation have been published [10]. They are usually divided into two categories: bottom-up and top-down approaches. As a typical bottom-up approach, OpenPose [5] detected all the body joints in the first stage, then associated them with person instances in the second stage. OpenPose [5] used a non-parametric representation called Part Affinity Fields (PAFs) with a greedy algorithm to generate the person instance.

On the other hand, top-down approaches reported better performance, where they first detected the bounding boxes of persons in the input image, then estimated the joint locations in each bounding box. Fang et al.[11] noticed that single-person pose estimation is sensitive to human detection. To solve this problem, they employed Symmetric Spatial Transformer Network (SSTN) in parallel with Single-Person Pose Estimator (SPPE) to extract a high-quality single-person region. Mask R-CNN [14] simultaneously predicted bounding boxes and body joints, which made the detection faster by sharing the features. Moreover, the new RoI alignment method enabled more accurate feature cropping. Chen et al.[7] proposed a network structure called Cascaded Pyramid Network (CPN), which consisted of two parts, GlobalNet and RefineNet. The former extracted a good feature representation, while the latter was employed to address the "hard" examples. The latest technology of HRNet [28] proposed an architecture that preserves high-resolution feature maps, which has been shown to be highly beneficial in multi-person pose estimation tasks. HRNet [28] consisted of multiple branches with different resolutions. Lower resolution branches captured contextual information and higher resolution branches preserved spatial information. With multi-scale fusions between branches, HRNet [28] can generate high resolution feature maps with rich semantic content.

Note that in all the papers mentioned above a heatmap is generated for each joint, in which the pixel value indicates the joint existence probability at that location. Zhang et al. [31] regarded heatmap as the de facto standard coordinate representation in human pose estimation.

2.2 Exploiting Temporal Information in Videos

For videos, a big challenge is how to exploit their temporal information [17]. Several prior methods [16], [17] tackled the video pose estimation task as a two-stage problem, first detecting the body joints in individual frames, and then applying temporal smoothing techniques. Later, recurrent networks especially LSTM [15] and GRU [8] were proposed for pose estimation [3], [20]. Moreover, 3D convolution is also useful for temporal information [13], [33]. Girdhar et al. [13] extended Mask-RCNN with 3D convolution for human pose estimation.

As demonstrated in other fields like action recognition [26], optical flow is a powerful source for temporal information because it explicitly contains motion information. In human pose estimation, optical flow was often used to temporally warp the heatmaps from another frame to the current frame [23], [27]. Song et al. [27] computed a dense optical flow between neighboring frames to propagate joint location estimates through time, and a flow based warping layer aligned the heatmaps to the current frame.

Recently, heatmap prediction/warping was realized by designing a particular subnet [4], [29]. PoseWarper [4] proposed convolutional layers with different dilation rates and deformable convolutions [9] to warp the heatmap from one frame to another. Note that the backbone network used in PoseWarper [4] was HRNet [28]. By using multiple frames, PoseWarper [4] was shown to be a promising approach to solve the challenging occlusion problem in human pose estimation and won the PoseTrack2017 Challenge 2 "Multi-frame Person Pose Estimation" [1]. However, the question of how much accuracy gain we can obtain by using an additional frame remains unanswered.

3. Preparations for Analysis

In this section, we briefly introduce the algorithm of our analysis target, PoseWarper [4], plus the basic informa-

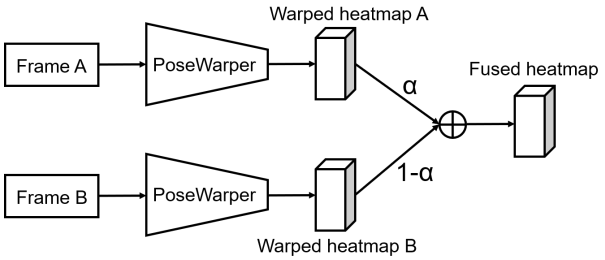


Fig. 1: The pipeline of PoseWarper [4]: two frames are used for human pose estimation in videos. With PoseWarper, a warped heatmap is generated from each frame and then a fused heatmap is computed by averaging two warped heatmaps.

tion on the PoseTrack dataset [17]. Also, we present our implementation details and experimental results from the PoseTrack dataset.

3.1 Analysis Target: PoseWarper

Figure 1 shows the pipeline of the cutting-edge technology, PoseWarper [4], which was selected as our analysis target due to its high performance. The model was trained on the training set of PoseTrack2017 with a pretrained model of HRNet on the COCO dataset. For simplicity, but without loss of generality, we used two frames during inference as mentioned in Section 1, which is different from the original paper.

As shown in Figure 1, given Frame A denoted as $\boldsymbol{I}(A)$ (3x384x288) and Frame B denoted as $\boldsymbol{I}(B)$, PoseWarper outputs two warped heatmaps $\boldsymbol{H}(A)$ (17x96x72) and $\boldsymbol{H}(B)$, which can be computed by

$$\boldsymbol{H}(A) = \boldsymbol{f}(\boldsymbol{I}(A); \boldsymbol{W}) \quad (1)$$

$$\boldsymbol{H}(B) = \boldsymbol{f}(\boldsymbol{I}(B); \boldsymbol{W}) \quad (2)$$

where $\boldsymbol{f}$ is the trained network with parameter of $\boldsymbol{W}$.

Then, we can fuse them together by simple averaging as

$$\boldsymbol{H}(A, B) = \alpha * \boldsymbol{H}(A) + (1 - \alpha) * \boldsymbol{H}(B) \quad (3)$$

where $\boldsymbol{H}(A, B)$ is the fused heatmap from Frames A and B, and α is a weight (set as 0.5 in our experiment).

Thus, the joint locations are estimated from the locations of maximum value on heatmaps.

$$\boldsymbol{P} = \arg\max \boldsymbol{H} \quad (4)$$

where $\boldsymbol{P}$ is the locations of each of the 17 joints, and $\boldsymbol{H}$ is a heatmap, which can be the fused heatmap $\boldsymbol{H}(A, B)$ or any warped heatmap $\boldsymbol{H}(A)/\boldsymbol{H}(B)$.

3.2 PoseTrack Dataset

PoseTrack dataset was released by the Max Planck Institute for Informatics and University of Bonn [17]. The videos in the dataset are from the MPII Human Pose dataset [2]. Currently, PoseTrack is one of the largest video datasets that include annotation for full set of body joints (17 joints in total) [10], [17], and includes 514 videos comprising 66,374 frames. The annotation is almost consistent with the MSCOCO format (except for the joints in the head), which is a dataset for human pose estimation in still images [19].

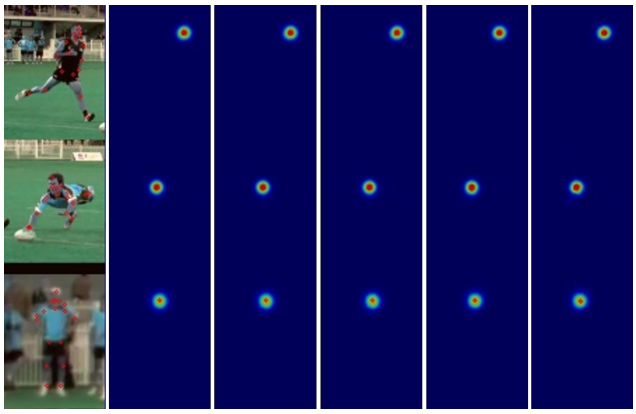


Fig. 2: Heatmaps of "nose" generated from five frames respectively. From left to right: sample input images with detected joints, heatmaps of "nose" from Frame t-2, Frame t-1, Frame t, Frame t+1, and Frame t+2.This figure shows that the heatmaps are successfully warped from other frames with rather correct location of "nose", which help pose estimation improve the accuracy by using multiple frames.

In commonly used datasets such as PoseTrack, MPII, and MSCOCO, the mean Average Precision (mAP) is generally used as a metric for the accuracy of human pose estimation [10], [17]. Using a defined threshold for the acceptable distance between estimated and actual joint location, each detection within this threshold is treated as a true-positive. The ratio of such true-positives to all detections is mAP.

3.3 Implementation Details and Results

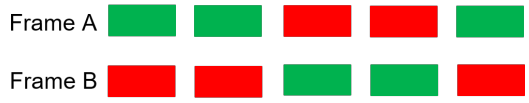
The source codes of PoseWarper [4] are available on the Internet[*1] and were used directly in our experiments. Most hyper-parameters are set as they were in the original paper during training. Table 1 shows the inference results from using just one frame in the validation dataset of PoseTrack 2017. Compared to the results of original paper, our results are a little lower (-0.2 in mAP), which may come from the randomness in training process. Therefore, we used the same trained model in our experiments to avoid this randomness problem. Compared to the baseline of HRNet [28] and the original PoseWarper [4] that used five frames, even using one frame in PoseWarper provides a high accuracy of pose estimation on the validation dataset of PoseTrack2017. Note that except for the current frame, the accuracy is a little worse than the baseline, which implies that the warped heatmaps from other frames are not perfect. However, as shown in Figure 2, the warped heatmaps from other frames have great potential in improving the accuracy of pose estimation by using multiple frames.

[*1] https://github.com/facebookresearch/PoseWarper

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean	gain1	gain2
Baseline	81.6	87.9	83.0	76.4	81.0	79.4	72.7	80.4	0	-0.6
PoseWarper	81.8	88.5	83.7	77.2	82.1	80.0	73.4	81.0	0.6	0
Frame t	81.6	88.1	83.1	76.5	81.7	79.6	73.0	80.6	0.2	-0.4
Frame t-1	81.2	88.0	82.8	75.9	81.4	79.2	72.5	80.2	-0.2	-0.8
Frame t+1	81.4	87.9	82.8	75.9	81.2	79.2	72.4	80.2	-0.2	-0.8
Frame t-2	80.4	87.2	81.4	74.3	80.5	78.0	71.0	79.1	-1.3	-1.9
Frame t+2	80.6	86.9	81.4	74.2	80.2	78.0	71.0	79.0	-1.4	-2.0

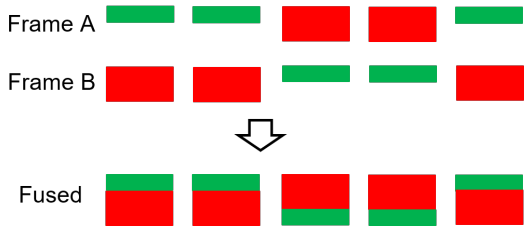
Table 1: Even using one frame in PoseWarper yields high accuracy of pose estimation on the validation dataset of PoseTrack2017. gain1: the accuracy gain from the baseline of HRNet [28]; gain2: the accuracy gain from original PoseWarper [4] that uses a total of five frames.

(a) Heatmaps from two frames (Frame A and Frame B) are not complementary

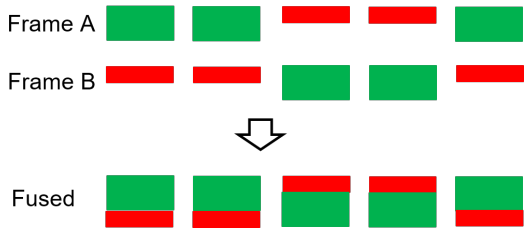


(b) heatmaps from two frames (Frame A and Frame B) are complementary

Fig. 3: There is no guarantee that two frames will yield better performance than a single frame. This figure illustrates two important cases given five test images: (a) ; (b) . Green: correct heatmaps; red: incorrect heatmaps. Note that Frame A or B may be any one in Table 1.



(a) Worse accuracy than single frame due to improper fusion



(b) Better accuracy than single frame due to proper fusion

Fig. 4: There is no guarantee that two frames will yield better performance than a single frame. This figure illustrates two different fusion strategies. Height denotes the fusion weight.

4. Analysis of Experimental Results

In this section, we report our findings by analyzing the experimental results, which answers two questions listed in Section 1. Namely, Section 4.1 is a qualitative analysis to obtain the conditions for performance improvement by using two frames. Section 4.2 is a quantitative analysis how much improvement is.

4.1 Key Observations for Qualitative Analysis

As mentioned before, suppose that we use two frames for human pose estimation in videos. For different test images, there are correct heatmaps[*2] and incorrect heatmaps of a particular joint as shown in Figure 3. If we want to produce a better performance (= more correct heatmaps) by using two frames rather than a single frame, it is necessary to have correct heatmaps from one frame for some test images and correct heatmaps from another frame for other test images, which is defined as the heatmaps from the two frames are complementary as shown in Figure 3(b). Otherwise, as shown in Figure 3(a), where the heatmaps are not complementary, it is impossible for the fused results to be better than Frame A alone. This is because there is no way to get correct heatmaps for the third and fourth test images in Figure 3(a). In a word, the first necessary condition for better performance by using two frames is that the frames should be complementary

Generally speaking, it is important to fuse the heatmaps from two frames. Figure 4 shows the concept of the fusion strategy. When a correct heatmap and an incorrect heatmap are generated from two frames respectively, it is necessary to give more weight to the correct heatmap than the incorrect one so that the fused heatmap is correct as shown in Figure 4(b). Otherwise, the fused results are even worse than a single frame as shown in Figure 4(a). In a word, the second necessary condition for better performance by using two frames is that a proper fusion strategy should be designed. This requires that we know which heatmap is correct in our fusion strategy.

Fortunately, we have confidence level in heatmaps. The maximum value in a heatmap is generally used to determine the degree of confidence. Usually, an incorrect heatmap has a lower maximum value (or confidence level) than a correct heatmap, which means we can tell which heatmap is correct in many cases. Therefore, as shown in Figure 5, a simple average works well, which was also demonstrated in PoseWarper [4]. Note that because the confidence level is not perfect, it is still useful in designing a smarter fusion strategy.

Finally, we conducted a comprehensive experiment by

[*2] Correct heatmap means the joint location estimated from the heatmap is true-positive.

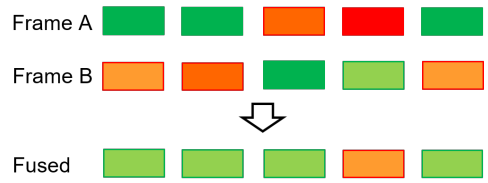


Fig. 5: We observe that even a very simple fusion strategy like averaging still works in human pose estimation thanks to the confidence effect. Color thickness denotes the confidence.

combining two frames to demonstrate the findings outlined above. As shown in Table 2, we obtained worse performance (80.5) by using Frame t from HRNet and PoseWarper than by just using PoseWarper (80.6), which infers no complementary information is available from two algorithms if we use the same frame. However, we do see a gain in accuracy if the two frames come from different frames, which are complementary.

4.2 Regression for Quantitative Analysis

For regression analysis, it is necessary to define the proper variables. Because we want to know the magnitude of the accuracy gain, it seems natural to use mAP as the output variable. However, there are two mAPs available from two frames. Therefore, we used the average of the mAPs from the two frames (denoted as $\overline{mAP}$ in Equation 5) and defined the difference from the accuracy of the fused heatmap as the output variable $\boldsymbol{y}$.

$$\overline{mAP} = \frac{mAP(\boldsymbol{H}(A)) + mAP(\boldsymbol{H}(B))}{2} \quad (5)$$

$$\boldsymbol{y} = mAP(\boldsymbol{H}(A, B)) - \overline{mAP} \quad (6)$$

where $mAP(\boldsymbol{H})$ denotes the accuracy from heatmap $\boldsymbol{H}$. Note that the output variable $\boldsymbol{y}$, which is called accuracy gain, is actually a relative value instead of an absolute value. This means the absolute performance depends not only $\boldsymbol{y}$ but also $\overline{mAP}$.

For the input variable, it seems intuitive to include an index to show the degree to which the two frames are complementary. In this work, we used the time difference as the index.

$$\boldsymbol{x} = \|A - B\| \quad (7)$$

where A and B denote the frame indexes of the two frames.

We calculated $\boldsymbol{x}$ and $\boldsymbol{y}$ using the experimental data from Tables 1 and 2 with Equations 5, 6, and 7, whose results are shown in Table 3. As shown in Figure 6, we conducted a linear regression between $\boldsymbol{x}$ and $\boldsymbol{y}$, with the result:

$$\boldsymbol{y} = 0.3008\boldsymbol{x} + 0.183 \quad (8)$$

4.3 Discussion

Limitations: The linear function in Equation 8 shows that if two frames are far away from each other, the accuracy gain is large. However, this cannot be extended indefinitely. We can imagine that if the time difference

Method	Head	Shoul.	Elbow	Wrist	Hip	Knee	Ankle	Mean
B & t	81.6	88.0	83.0	76.4	81.5	79.5	72.9	80.5
t-2 & t-1	81.2	88.1	82.9	76.0	81.6	79.4	72.5	80.3
t-2 & t	81.6	88.3	83.3	76.5	82.0	79.8	73.0	80.7
t-2 & t+1	81.6	88.3	83.3	76.8	81.9	79.9	72.9	80.7
t-2 & t+2	81.4	88.1	83.0	76.3	81.6	79.5	72.4	80.4
t-1 & t	81.7	88.4	83.4	76.6	81.9	79.8	73.2	80.8
t-1 & t+1	81.8	88.4	83.5	77.0	82.0	79.9	73.3	80.9
t-1 & t+2	81.7	88.3	83.3	76.7	81.9	79.7	73.0	80.7
t & t+1	81.8	88.3	83.4	76.8	81.9	79.8	73.1	80.8
t & t+2	81.6	88.1	83.2	76.6	81.9	79.7	72.9	80.6
t+1&t+2	81.4	87.9	82.9	76.1	81.4	79.2	72.4	80.3

Table 2: Accuracy by using two frames in PoseWarper on the validation dataset of PoseTrack2017. The red row shows worse accuracy than a single frame.

Method	$\boldsymbol{x}$	$\overline{mAP}$	$mAP(\boldsymbol{H}(A, B))$	$\boldsymbol{y}$
Baseline & Frame t	0	80.471	80.5	0.029
Frame t-2 & Frame t-1	1	79.65	80.3	0.65
Frame t-2 & Frame t	2	79.85	80.7	0.85
Frame t-2 & Frame t+1	3	79.65	80.7	1.05
Frame t-2 & Frame t+2	4	79.05	80.4	1.35
Frame t-1 & Frame t	1	80.4	80.8	0.4
Frame t-1 & Frame t+1	2	80.2	80.9	0.7
Frame t-1 & Frame t+2	3	79.6	80.7	1.1
Frame t & Frame t+1	1	80.4	80.8	0.4
Frame t & Frame t+2	2	79.8	80.6	0.8
Frame t+1 & Frame t+2	1	79.6	80.3	0.7

Table 3: The input and output variables for linear regression, calculated from Tables 1 and 2.

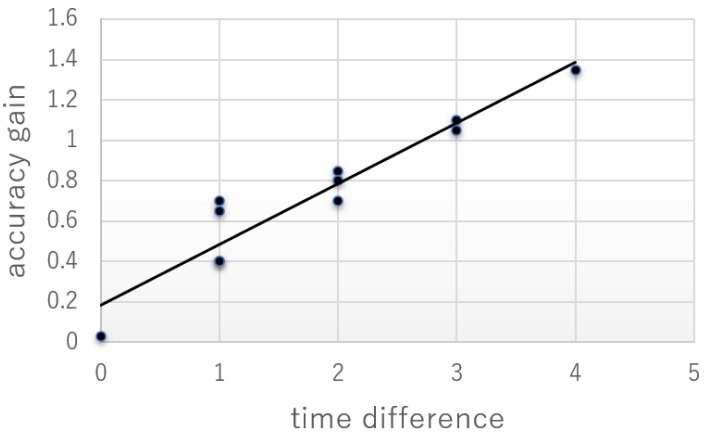


Fig. 6: The accuracy gain depends on the time difference of two frames, which forms a linear function.

between two frames is too large, the accuracy gain will be saturated because the information in a faraway frame is not related to the current frame, resulting in difficulty in warping heatmaps. Actually, our experiment is limited to the short-term (less than five frames).

5. Conclusions

As far as we know, this paper presents for the first time a quantitative analysis of multiple-frame effect for human pose estimation in videos. Interestingly, we observe that there is no guarantee that two frames will yield better performance than a single frame. Furthermore, we specify here two necessary conditions for performance improvement (1+1>1): the frames should be complementary, and a proper fusion strategy should be designed. In addition, by analyzing the experimental results of PoseWarper [4] in PoseTrack2017 [17], we demonstrate that a linear regression works well to model the relationship between the accuracy gain and time difference of two frames.

References

[1] : (online), available from ⟨https://posetrack.net/leaderboard.php⟩.
[2] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B.: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014).
[3] Artacho, B. and Savakis, A.: UniPose: Unified Human Pose Estimation in Single Images and Videos, arXiv preprint arXiv:2001.08095 (2020).
[4] Bertasius, G., Feichtenhofer, C., Tran, D., Shi, J. and Torresani, L.: Learning Temporal Pose Estimation from Sparsely-Labeled Videos, Advances in Neural Information Processing Systems 32 (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., eds.), Curran Associates, Inc., pp. 3027–3038 (online), available from ⟨http://papers.nips.cc/paper/8567-learning-temporal-pose-estimation-from-sparsely-labeled-videos.pdf⟩ (2019).
[5] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S. and Sheikh, Y. A.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).
[6] Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794 (2016).
[7] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G. and Sun, J.: Cascaded Pyramid Network for Multi-Person Pose Estimation (2018).
[8] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).
[9] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H. and Wei, Y.: Deformable convolutional networks, Proceedings of the IEEE international conference on computer vision, pp. 764–773 (2017).
[10] Dang, Q., Yin, J., Wang, B. and Zheng, W.: Deep learning based 2D human pose estimation: A survey, Tsinghua Science and Technology, Vol. 24, No. 6, pp. 663–676 (2019).
[11] Fang, H.-S., Xie, S., Tai, Y.-W. and Lu, C.: RMPE: Regional Multi-person Pose Estimation, ICCV (2017).
[12] Freund, Y. and Schapire, R. E.: A desicion-theoretic generalization of on-line learning and an application to boosting, European conference on computational learning theory, Springer, pp. 23–37 (1995).
[13] Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M. and Tran, D.: Detect-and-track: Efficient pose estimation in videos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 350–359 (2018).
[14] He, K., Gkioxari, G., Dollár, P. and Girshick, R.: Mask r-cnn, Proceedings of the IEEE international conference on computer vision, pp. 2961–2969 (2017).
[15] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, Vol. 9, No. 8, pp. 1735–1780 (1997).
[16] Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B. and Schiele, B.: Arttrack: Articulated multi-person tracking in the wild, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6457–6465 (2017).
[17] Iqbal, U., Milan, A. and Gall, J.: Posetrack: Joint multi-person pose estimation and tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2011–2020 (2017).
[18] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, pp. 3146–3154 (2017).
[19] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft coco: Common objects in context, European conference on computer vision, Springer, pp. 740–755 (2014).
[20] Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J. and Lin, L.: Lstm pose machines, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5207–5215 (2018).
[21] Marechal, C., Mikołajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C. and Węgrzyn-Wolska, K.: Survey on AI-Based Multimodal Methods for Emotion Detection, High-Performance Modelling and Simulation for Big Data Applications, Springer, pp. 307–324 (2019).
[22] Park, E., Han, X., Berg, T. L. and Berg, A. C.: Combining multiple sources of knowledge in deep cnns for action recognition, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 1–8 (2016).
[23] Pfister, T., Charles, J. and Zisserman, A.: Flowing convnets for human pose estimation in videos, Proceedings of the IEEE International Conference on Computer Vision, pp. 1913–1921 (2015).
[24] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. and Gulin, A.: CatBoost: unbiased boosting with categorical features, Advances in neural information processing systems, pp. 6638–6648 (2018).
[25] Schapire, R. E.: The strength of weak learnability, Machine learning, Vol. 5, No. 2, pp. 197–227 (1990).
[26] Simonyan, K. and Zisserman, A.: Two-stream convolutional networks for action recognition in videos, Advances in neural information processing systems, pp. 568–576 (2014).
[27] Song, J., Wang, L., Van Gool, L. and Hilliges, O.: Thin-slicing network: A deep structured model for pose estimation in videos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4220–4229 (2017).
[28] Sun, K., Xiao, B., Liu, D. and Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, CVPR (2019).
[29] Wang, J., Qiu, K., Peng, H., Fu, J. and Zhu, J.: AI Coach: Deep Human Pose Estimation and Analysis for Personalized Athletic Training Assistance, Proceedings of the 27th ACM International Conference on Multimedia, pp. 374–382 (2019).
[30] Zhang, D., Guo, G., Huang, D. and Han, J.: Poseflow: A deep motion representation for understanding human behaviors in videos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6762–6770 (2018).
[31] Zhang, F., Zhu, X., Dai, H., Ye, M. and Zhu, C.: Distribution-Aware Coordinate Representation for Human Pose Estimation, Proceedings of the IEEE conference on computer vision and pattern recognition (2020).
[32] Zhou, H., Mian, A., Wei, L., Creighton, D., Hossny, M. and Nahavandi, S.: Recent advances on singlemodal and multimodal face recognition: a survey, IEEE Transactions on Human-Machine Systems, Vol. 44, No. 6, pp. 701–716 (2014).
[33] Zhou, L., Chen, Y., Wang, J. and Lu, H.: Progressive Bi-C3D Pose Grammar for Human Pose Estimation, Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (2020).