NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY
[IJIERT] ISSN: 2394-3696 Website: ijiert.org
VOLUME 7, ISSUE 12, Dec.-2020

DEEP LEARNING BASED HUMAN POSE ESTIMATION USING OPENCV

Jupalle Hruthika
Electronics Engineering Sardar Vallabhbhai National Institute of Technology
Country—-India
hruthikareddy8@gmail.com

Pulipati Krishna Chaitanya
Mechanical Engineering Sardar Vallabhbhai National Institute of Technology
Country: India
Krishnachaitanya887@gmail.com

Goli shiva Chaithanya
Computer Science KLUniversity Country -India
chaithushiva2345@gmail.com

Abstract
In vision-based human activity analysis, human pose estimation is an important study area. The goal of human pose estimation is to estimate the positions of the human articulation joints in 2D/3D space from photographs or movies. Because of the complication ofreal-world settings and a wide range of human stances, vision-based human poses. Estimation is a difficult task. Deep learning's rapid advancement has recently attracted a lot of attention. The simulation of the processing and reasoning capacities of the human brain has received a lot of attention. The visual system of humans. As a result, it is critical to continue to investigate. Deep learning techniques are used to estimate human pose based on imagery. a video-based 2D pose estimation approach that incorporates a multi-scale TCE module into the encoder-decoder network design to explore temporal consistency in videos explicitly. At the feature level, the TCE module uses the learnable offset field to capture the geometric transition between neighbouring frames. We further investigate multi-scale geometric changes at the feature level by incorporating the spatial pyramid into the TCE module, which results in even more performance gains.

Keywords: Human, Pose, Deep, Learning, Vision

INTRODUCTION

HUMAN posture estimation (HPE), which has received a lot of attention in the computer vision field, entails predicting the configuration of human body parts using sensor input data, such as photographs and videos. HPE gives geometry and motion information on the human body, which has been used in a variety of applications (for example, human-computer interaction, motion analysis, augmented reality (AR), virtual reality (VR), healthcare, and so on). Deep learning solutions have been demonstrated to outperform traditional computer vision approaches in a variety of tasks, including picture classification, thanks to the rapid growth of deep learning solutions in recent years. However, obstacles such as occlusion, insufficient training data, and depth ambiguity must to be solved. 2D HPE from photos and movies with 2D pose annotations is simple to achieve, and good performance has been achieved for a single person's human pose estimation using deep learning algorithms. In controlled lab contexts, motion capture devices can acquire 3D pose annotation; nevertheless, they have limitations in real-world situations. The fundamental issue in 3D HPE from monocular RGB photos and videos is depth ambiguity. The key issue that needs to be addressed in multiview setups is viewpoint affiliation. Some studies have used sensors such as depth sensors, inertial measurement units (IMUs), and radio frequency devices, although these methods are usually expensive and necessitate specialised gear. Human posture estimation is also utilised in video surveillance, human-computer interface, sports analysis, virtual reality, animation development, and other domains. Human pose estimation, for example, can be used to track human subjects' mobility in interactive gaming. Microsoft's Kinect, for example, popularised the use of 3D pose estimation to track the motion of

the human player and render the activity of the virtual character. Human posture estimate can rebuild the athlete's motion from daily training recordings in sports analysis. CGI applications can also benefit from human pose estimation. If their human position can be calculated, graphics, styles, fancy improvements, equipment, and artwork can be superimposed. The produced images can naturally match the person as he or she walks by tracking the fluctuations of this human stance.

2D human posture estimation and 3D human pose estimation are the two types of human pose estimation. Both 2D and 3D human posture estimation are difficult jobs due to the complexity of the real world and the diversity of human stances. Deep learning has been extensively used on the job of human pose estimate in recent years, thanks to the rapid growth of Convolutional Neural Networks (CNNs). Despite the fact that deep learning-based algorithms have made great progress, they nevertheless face several obstacles. The majority of available approaches for 2D human posture estimation focus on building novel network topologies for image-based 2D pose estimation. movements. Although these methods can be applied directly to video data, they frequently produce poor results because image-based methods cannot take use of the rich temporal information contained in video data. A typical neural network model for 3D human pose estimation requires a substantial amount of training data. Annotating 3D human joint locations, on the other hand, is a time-consuming operation. Furthermore, there are solid geometrical theories for projecting 2D images onto 3D skeletons. Using a neural network to approximate this projection alone could result in the network being overfitted with training data.

BACKGROUND INFORMATION

Following are some key work details which were discussed by keeping the related thing in the investigation. The goal of human pose estimate, sometimes called human keypoint estimation,

is to find anatomical keypoints in the human body. It's a crucial task in the subject of computer vision, with substantial theoretical implications and widespread applications in disciplines like human activity recognition, sports analysis, and human-computer interaction. Many world- class research teams and institutes have recently committed significant resources to studying this issue.

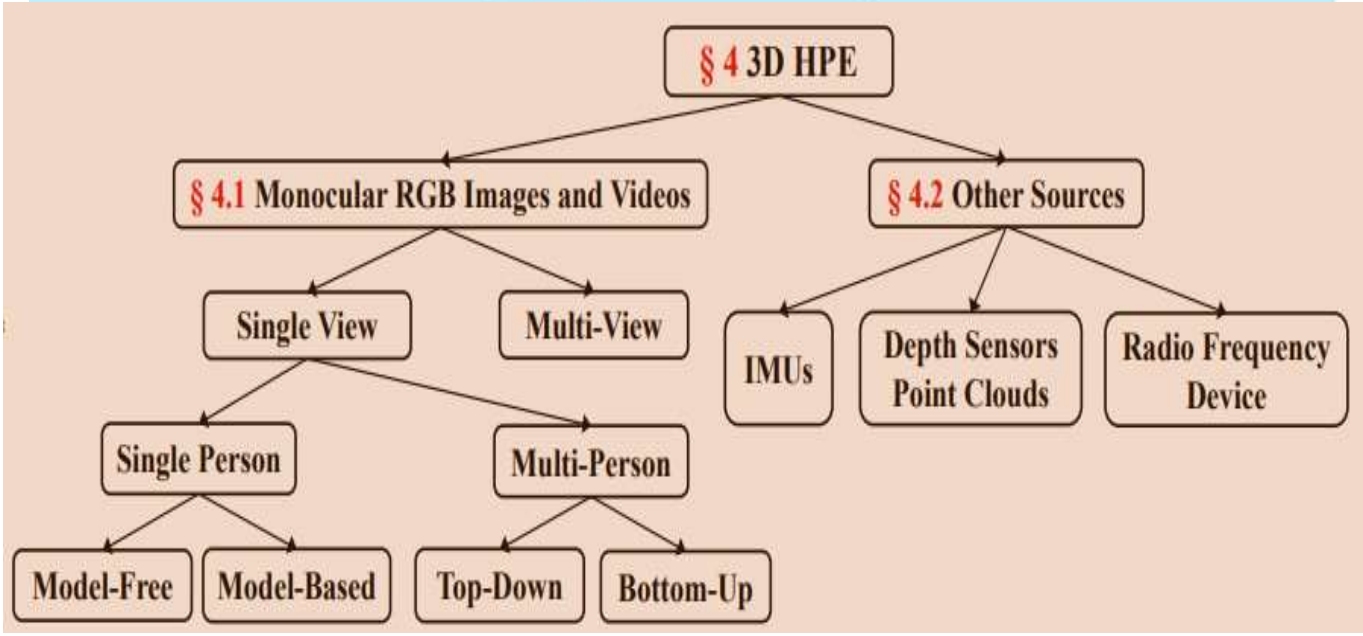

Fig. 1 A hierarchy of literature survey

The Robotics Institute at Carnegie Mellon University, for example, has developed Panoptic Studio, a large-scale multi-view human motion capture system. Common Objects in Setting (COCO) is a large-scale 2D human pose dataset created by Microsoft Research that collects photos of complicated everyday scenarios incorporating common objects in their natural context. Simultaneously, they held competitions and workshops that aided in the development of 2D pose estimation technology. The MPII and MPI-INF-3DHP

datasets, which are extensively used 2D and 3D human pose datasets, were also proposed by the Max Planck Institute for Computer Science. In top computer vision conferences and journals, such as CVPR (IEEE Conference on Computer Vision and Pattern Recognition), ICCV (International Conference on Computer Vision), ECCV (European Conference on Computer Vision), PAMI (IEEE Transaction on Pattern Analysis and Machine Intelligence), TIP (IEEE Transaction on Image Analysis), and IJCV (International Conference on Computer Vision), human pose estimation has already become one of the hottest topics. Human pose estimation is separated into two types: 2D pose estimation and 3D pose estimation, which estimate the positions of human joints in two-dimensional and three-dimensional space, respectivelyThe majority of traditional image-based 2D posture estimate algorithms are part-based from the bottom up. These methods treat the human posture as a collection of human body parts and use the deformable model to characterise body component spatial connections. The pictorial structure model for visual object representations was proposed by Fishler et al. [25] in 1973. Following that, Felzenszwalb et al.The poselet prior was utilised by Pishchulin et al. to improve the visual structure model. These methods rely on handcrafted characteristics to detect human body components, such as the Histogram of Oriented Gradient (HoG) and Scale-Invariant Feature Transform (SIFT), and then utilise a dynamic programming algorithm to get the best human pose configuration. However, for photos of complicated everyday scenarios with truncated or badly occluded human joints, these techniques lack generalisation ability. Researchers attempted to apply Convolutional Neural Networks (CNNs) to human posture estimation after seeing the success of deep learning in object classification and detection. Meanwhile, large-scale human pose datasets such as FLIC [73], MPII [1], and Microsoft COCO [53] are available, allowing deep networks to be trained. Because stacked convolution and pooling layers allow CNNs to learn high-level visual characteristics, these approaches can directly predict human joint locations from input photos. Recent research aims to improve video-based posture estimation performance by incorporating temporal information into sophisticated deep models. The most frequent methods [67, 11, 78, 97] use optical flow to explore temporal context.

Because optical flow describes the distribution of apparent movement velocities, it can help refine the projected heatmaps by capturing geometric transitions between frames. Song et al. [78], for example, employed optical flow to take use of visual evidence from neighbouring frames. To improve the performance of video pose estimation, Pfister et al. [67] used optical flow to align output heatmaps from surrounding frames. Other approaches [28, 54] use Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), to capture temporal dependency. RNNs have become dominant tools for sequence tasks due to their power in long-range temporal representation. Luo et al., for example, suggested a recurrent model with LSTM to consider temporal information for video posture estimation. Gkioxari et al. proposed a CNN-based chained model in which the pose prediction is based on both the input and the output of the preceding frame. There are also approaches for learning representations of video clips using 3D convolution. Girdhar et al. [27] increased the Mask R-2D

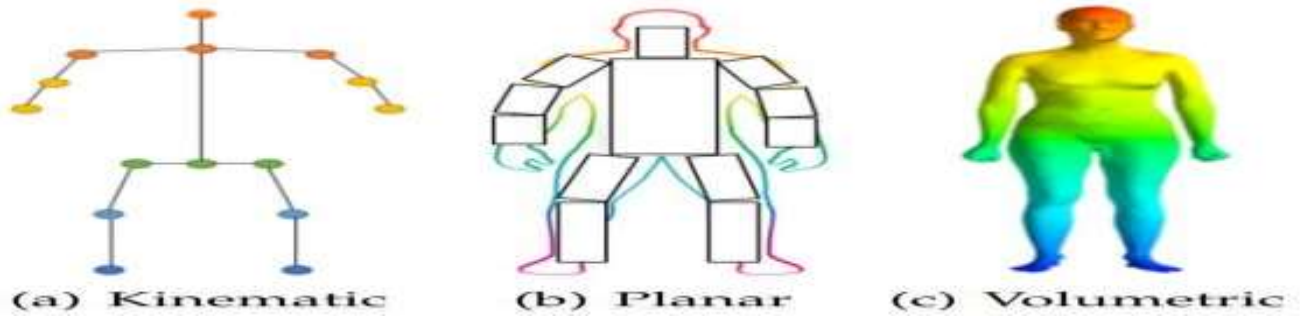

Fig. 2 Type of human body

CNN's convolution to 3D, allowing it to use temporal information from video clips to generate more accurate position predictions in videos.

MONOCULAR 3D POSE ESTIMATION

The huge potential of 3D human posture estimation in diverse applications such as human- computer interaction, virtual reality, and action detection has piqued curiosity. Many researchers [55, 65] used a neural network to predict 3D human poses from monocular photos as a result of deep learning's success.

There are two key issues in employing neural networks to estimate 3D postures. A conventional neural network model, for starters, necessitates a substantial amount of training data. 3D Marker-based Motion Capture is used to collect pose annotations (MoCap) system, which is a time-consuming procedure. In this research, we offer a unique self-supervised strategy for training a 3D posture estimation model that takes advantage of the geometric prior. We define 3D pose estimation as a combination of 2D keypoint estimate and 2D-to-3D pose lifting. Our work focuses on training the 2D-to-3D lifting network without utilising any additional 3D ground-truth data, and the first stage is compatible with any state-of-the-art 2D keypoint detector. We construct the transform re-projection loss in particular to tackle the depth ambiguity problem

RESULT AND ANALYSIS

The training approach has two stages in order for the proposed two-branch network to converge without explicit 3D pose monitoring. First, we use the Lpre-train loss to pre-train the network. We train the network for 20 epoches with a learning rate of 0.001 using Adam as the optimizer. The network is then trained for 300 epoches using the LT loss. The rate of learning begins at and decreases by 0.1 per 100 epoches. During the assessment, we solely use the 2D-to- 3D lifting branch to predict the relative 3D poses in the camera space, rather than the root position branch, to maintain compatibility with other works. Pytorch, a deep learning toolbox, is used to develop our technique. We compare the suggested transform re-projection loss to an existing popular technique, adversarial loss, in order to assess its effectiveness. On the H36M dataset, we create many versions and compare the outcomes under Protocol #1 (MPJPE) and Protocol #2 (P-MPJPE). As inputs, all variations use 2D postures retrieved by the CPN network. The quantitative results are presented in Table, and Figure depicts the outcomes of several versions on multiple hard samples, such as with severe self-occlusion or far from the camera.

ANALYSIS OF NETWORK PERTAINING

This point details about the pre training of the data set which is very relevant for our analysis.

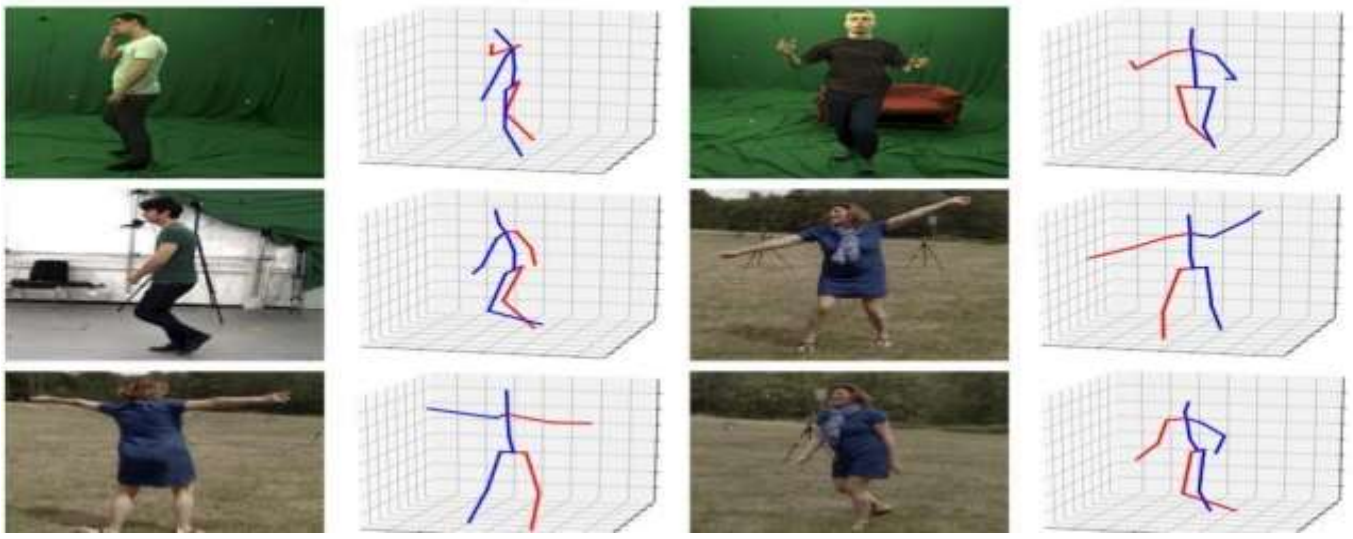

Fig. 3 Training dataset of 3DHP

ANALYSIS OF CONSISTENT FRAME WORK

In this section, we look at how successful the suggested consistent factorization loss is. All variations are trained on the Human3.6M train set, and Table 4.1 shows their per- action P-MPJPE on the Human3.6M test set. Clearly, our strategy performs the best out of all the alternatives. The pre-trained hierarchical lexicon aids in obtaining superior outcomes as compared to the baseline. However, if only the hierarchical dictionary is used, the benefit is minimal.

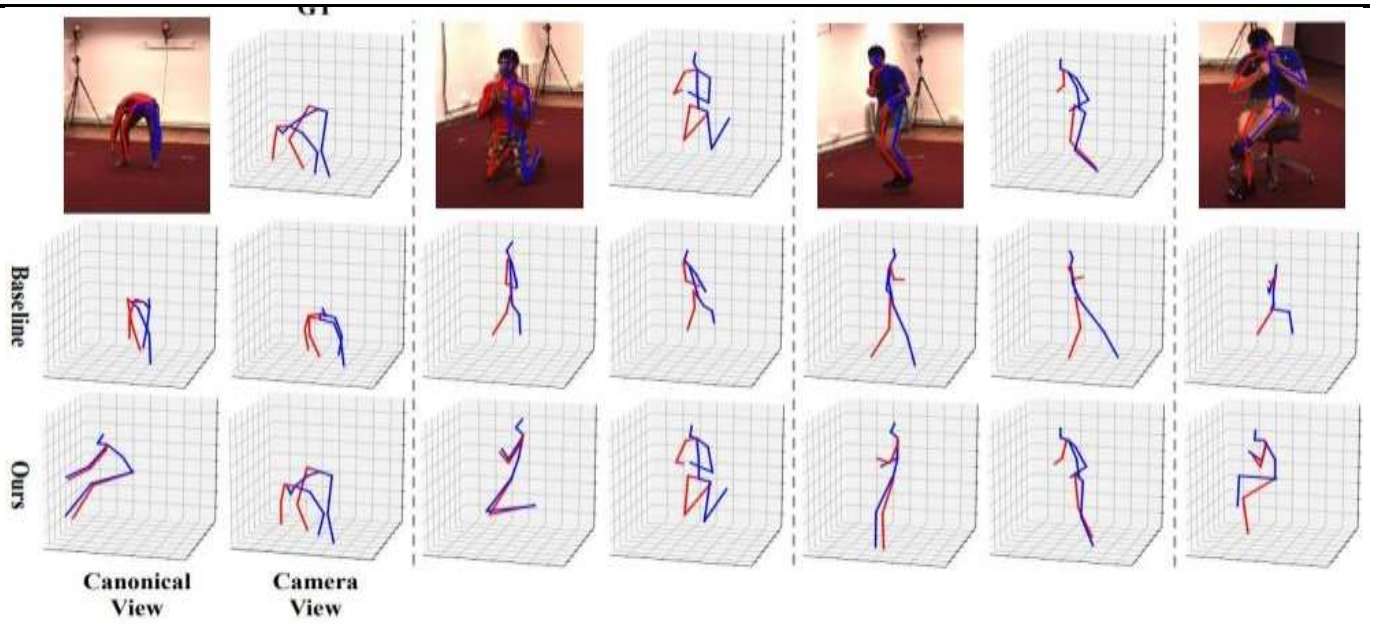

Fig. 4 Comparison with baseline

ANALYSIS OF HIERARCHICAL DICTIONARY

The efficacy of the hierarchical dictionary is examined in this paper. On the Human3.6M, it shows comparisons with modern dictionary-based approaches. AIGN [89] uses PCA to learn a 3D pose vocabulary and adds adversarial loss as a restriction. C3DPO [60] employs a single-level vocabulary that is learned in tandem with a 3D pose estimate network. Distill [91] is a weakly-supervised technique for learning a 3D pose estimation network from a lexicon obtained via NRSfM. Our strategy, as indicated in Table 4.2, produces the best results of all. We build a

single-level dictionary, comparable to C3DPO, for additional comparisons. Ours-SD outperforms C3DPO with the consistent factorization restriction, achieving 85.8 vs. 95.6 (mm) MPJPE. Furthermore, the hierarchical dictionary aids in achieving better outcomes than the single-level dictionary, with MPJPE and P-MPJPE decreasing by 3.9 and 5.2 (mm) respectively.

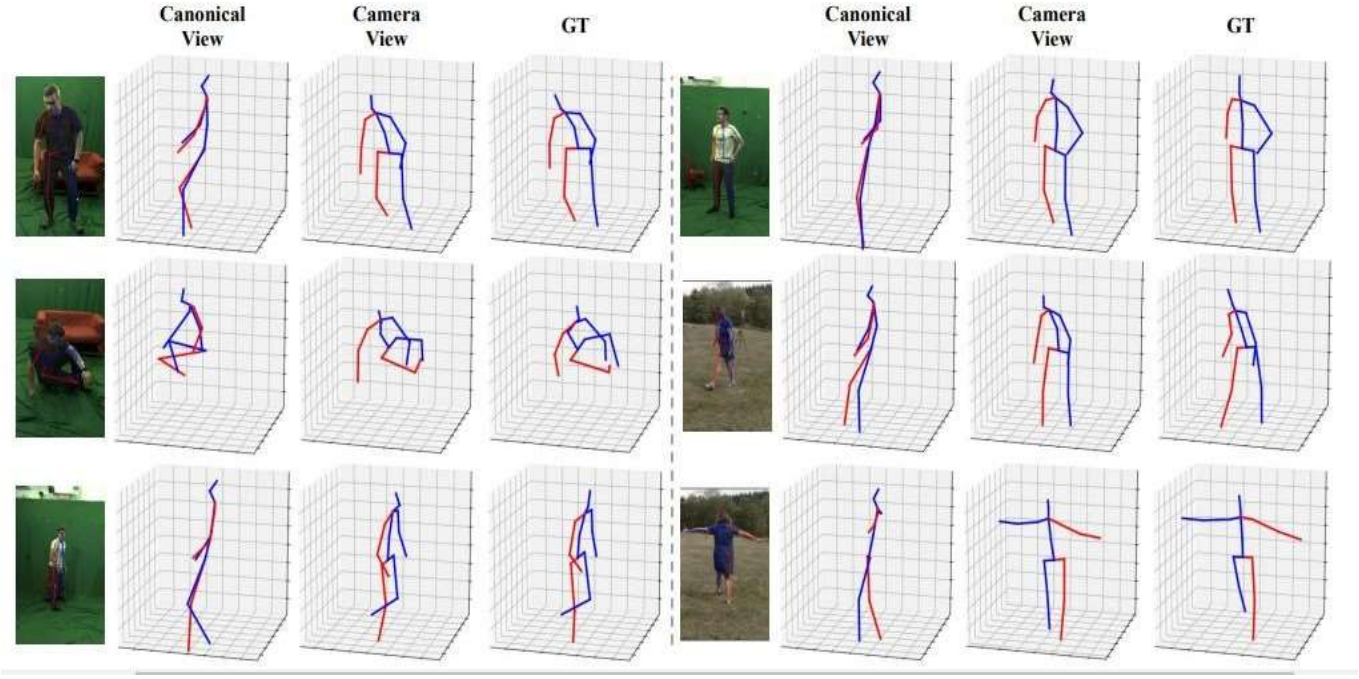

Fig. 5 Visualization analysis of result

COMPARISON OF RESULTS WITH OTHER RESEARCHERS

Following table describing the comparative description of this vestigation with other methods.

Table: 1 Comparison of results

Method	MPJPE	P-MPJPE
Wandt *et al.* CVPR'19	89.9	65.1
Zhou *et al.* ICCV'17	-	64.9
Drover *et al.* ECCV'18	-	64.6
Pavlakos *et al.* CVPR'17	118.4	-
Rhodin *et al.* ECCV'18	-	98.2
Chen *et al.* CVPR'19	-	68.0
Kocabas *et al.* CVPR'19	77.8	70.7
Tung *et al.* ICCV'17	97.2	-
Wang *el al.* ICCV'19	86.4	62.8
Novotny *et al.* ICCV'19	95.6	-
Ours	81.9	52.1

ANALYSIS OF GENERALIZATION ABILITY

To test the proposed model's generalization ability, we trained it on the Human3.6M dataset and tested it on the MPI-INF-3DHP dataset, which contains complex outdoor scenes. Figure illustrates some visualization results demonstrating that our technology can successfully recover 3D poses on datasets without having been trained on them. Furthermore, in this setup table, our approach can get 70.6 percent PCK3D and 36.6 percent AUC.

CONCLUSION AND FUTURE SCOPE

Conclusion

In the field of computer vision, human pose estimation is a hot issue of research. This thesis investigated deep learning-based 2D and 3D human position estimation and suggested a number of models, ranging from video-based 2D pose estimation to self-supervised 3D pose estimation. The following is a summary of the important innovative contributions:

For specifically exploring temporal consistency in films, we offer the multi-scale TCE module and embed it into the encoder-decoder network architecture. At the feature level, the TCE module uses the learnable offset field to capture the geometric transition between neighbouring frames. It can explicitly represent the temporal consistency information in an end-to-end network, unlike existing model-based techniques.

It is more computationally efficient than existing post-enhancement approaches since it does not involve additional optical flow computations. We further investigate multi- scale geometric changes at the feature level by incorporating the spatial pyramid into the TCE module, which results in even more performance gains.

A root position regression branch is also introduced to restore the global 3D poses during training. The network can save the scale information of re-projected 2D poses in this fashion, which improves the accuracy of predicted 3D poses. Furthermore, during training, this method just uses geometry information, resulting in improved generalization ability.

To solve the projection ambiguity problem, we offer the consistent factorization network, which entirely disentangles the 3D human shape and camera viewpoint. To this purpose, we create a simple and effective

loss function that constrains the canonical 3D human position using multi-view information. Furthermore, we characterize a 3D human pose as a combination of a dictionary of 3D pose base and use geometric information from 3D human poses to learn a hierarchical dictionary from 2D human poses by solving the NRSfM issue.

When compared to the single-level dictionary, the hierarchical dictionary can be learned without the use of 3D human posture annotations and has a greater expressive ability.

FUTURE SCOPE OF WORK

To improve the performance of 2D pose estimation in multi-person videos, we will aim to create a unified framework integrating the multi-scale TCE module with the multi- person tracking technique.

In terms of 3D posture estimation, we'll look into depth maps and point cloud data in the future. The cost of obtaining depth map and point cloud data will decrease as depth cameras and radar sensors become more widely available on mobile devices. Absolute depth information may be obtained from the depth map and point cloud, successfully resolving the projection ambiguity problem.

REFERENCES

1) Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NeurIPS, 2012.
2) J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in CVPR, 2015.
3) S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE TPAMI, 2016.
4) T. B. Moeslund and E. Granum, "A survey of computer visionbased human motion capture," CVIU, 2001.
5) T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances ¨ in vision-based human motion capture and analysis," CVIU, 2006.
6) R. Poppe, "Vision-based human motion analysis: An overview," CVIU, 2007.
7) X. Ji and H. Liu, "Advances in view-invariant human motion analysis: a review," IEEE TSMC, 2009.
8) M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," IEEE Journal of Selected Topics in Signal Processing, 2012.
9) Z. Liu, J. Zhu, J. Bu, and C. Chen, "A survey of human pose estimation: the body parts parsing based methods," JVCIR, 2015.
10) W. Gong, X. Zhang, J. Gonzalez, A. Sobral, T. Bouwmans, C. Tu, ` and E.-h. Zahzah, "Human pose estimation from monocular images: A comprehensive survey," Sensors, 2016.
11) N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," CVIU, 2016.
12) Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," CVIU, 2020.
13) T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation," IEEE Access, 2020.
14) E. Marinoiu, D. Papava, and C. Sminchisescu, "Pictorial human spaces: How well do humans perceive a 3d articulated pose?" in ICCV, 2013.
15) S. Zuffi, O. Freifeld, and M. J. Black, "From pictorial structures to deformable structures," in CVPR, 2012.
16) S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation." in BMVC, 2010.
17) Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in CVPR, 2017.
18) X. Chen and A. L. Yuille, "Parsing occluded people by flexible compositions," in CVPR, 2015.

19) D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," ACM TOG, 2017.
20) S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in FG, 199
21) 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, 2014, pp. 3686–3693.
22) M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1014–1021.
A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3d human pose estimation in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3395–3404.
23) S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," International Journal of Computer Vision, vol. 92, no. 1, pp. 1–31, 2011.
24) Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3d pictorial structures for multiple human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1669–1676.
25) V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2017, pp. 468–475.
26) E. Brau and H. Jiang, "3d human pose estimation via deep learning from 2d annotations," in Proceedings of the International Conference on 3D Vision, 2016, pp. 582–591.
27) T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 3, pp. 500–513, 2010.
28) Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.
29) J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3063–3072.
30) C.-H. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7035–7043.
31) L. Chen, H. Ai, R. Chen, Z. Zhuang, and S. Liu, "Cross-view tracking for multi-human 3d pose estimation at over 100 fps," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
32) X. Chen, K.-Y. Lin, W. Liu, C. Qian, and L. Lin, "Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 895–10 904.
33) Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103–7112.
A. Cherian, J. Mairal, K. Alahari, and C. Schmid, "Mixing body-part sequences for human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2361–2368