The Seybold Report

ISSN 1533-9211

DOI 10.5281/zenodo.7127542

DETECTING AND TRACKING: ASSESSMENT OF WELL-ORGANIZED POSES IN VIDEOS DYNAMICALLY

Dr. PARMANAND PRABHAT[1], MARNENI KOMALI[2], VELURU HARSHITHA[3], TATIKONDALA SHASHIDHAR[4] and NAGINENI SAI LASYA[5]

[1]Assistant Professor, Department of Computer Science & Engineering, Amrita Sai Institute & Technology, India
[2,3,4,5] Students, Department of Computer Science& Engineering, Vellore Institute of Technology, Amaravathi, India.

Abstract

Detection and tracking of human body key points in a multi-person video is the focus of this research article. In this, we use the most recent developments in video-based human-key point identification. Our technique uses key point estimate in frames or short video clips that include numerous people. Human position estimate and tracking is a newer method for locating a person's most important physical features. Human body language may be deciphered by computers using posture detection and tracking. It also aids in estimating the positions of human body parts and joints in photos and films, which is a huge benefit. It is Move Net, which uses temporal information from short video clips to anticipate quick and reliable outcomes that are utilized to estimate posture at the frame level. There are 17 critical points in an individual's movement that the Move Net model of motion estimation algorithm can identify. It is a highly quick and accurate model. Individual key points, and sometimes the affinities between them, are identified in a bottom-up model known as Move Net and then the predictions are aggregated into instances, which also employs heat maps to precisely locate key points on a human body. A feature extractor and a group of prediction heads make up this system's design. Multi-person video pose estimation, or MPII, is a technique we use to test different aspects of our model.

Keywords: occlusions, key points, pose estimation, Move Net

INTRODUCTION

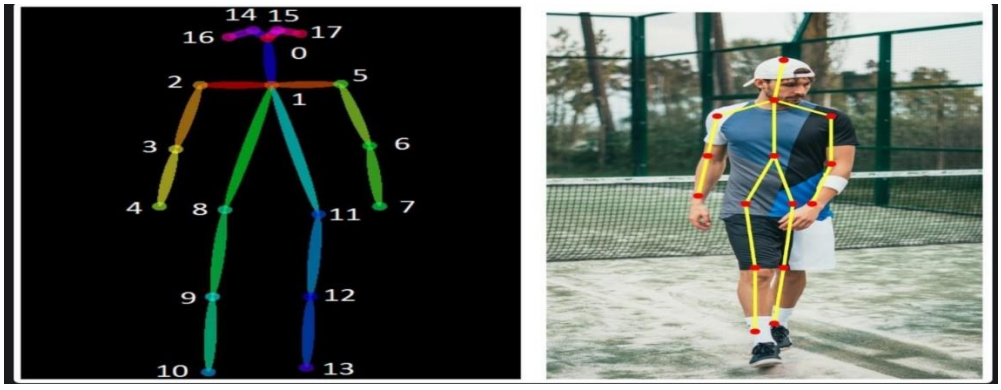
Recently, visual understanding like object recognition, has observed a deep visual representations. Understanding of human behavior in normal images has been the control center of visual tasks due to its significance in the practical application. Usually, person and pose detection and estimation from a single image have emerged as a leading and challenging visual recognition problem because it isa bit complex task to work on. Initially, single image understanding was a very difficult and complex task ,so later on video understanding made a slower improvement compared to image recognition i.e; they grew into more complex task compared to image recognition because a video is that which contains different number of frames . Combination of many frames is equal to video it may be a short video or long video.

In this, we point on the issue of human pose tracking in multi-person videos which tracks and estimates the pose of a person or human over time. In this there are many challenges which are occurred including the pose changes of a person throughout the video and also presence of multiple overlapping of instances over time. The tracking optimization is only managed for linking the frame-level predictions whereas the system has no capability to improve the location of key points on the frames. So, the key points are said to be imperfectly localized in a particular frame which leads to motion blur, occlusions also .To remove this limitation we use a model

ISSN 1533-9211

DOI 10.5281/zenodo.7127542

Move Net which is an ultra- fast and accurate estimator model which detects nearly 17 key points on the given frame i.e; on a human body which is especially designed for sports, exercises etc. It even helps in identifying the person who is 6 meters far away from the camera. In this model we have two variants such as thunder and lighting. Thundering Move Net variant is used when we need highest accuracy for the output and lighting is used where we want some accurate results. In this work, we will be using lighting variant which helps in detecting the key points more accurately and also can work on any application .It will detect key points with an accurate value and produces the required output. In this we will also implement GPU which is used to set the memory growth to limited in order to get rid of the memory error.



We provide training and evaluation for our approach on the MPII dataset which hold real time videos of people in different positions and scenes like when performing yoga, workouts, jogging and etc locates the key points in the frames and produces the result in the form of frames.

LITERATURE SURVEY

The title of the reference paper is Real time Multi-Person 2D Pose Estimation using Part Affinity Fields. The authors of this paper are Zhe Cao, Tomas Simon, Shih-En Wei,Yaser Sheikh. The methodology used in this is that it takes the complete image as input and part confidence maps and affinity fields and by using Bipartite and greedy bottom-up parsing technique, matching poses are tracked with high accuracy. The advantage of this work is that it detects the 2D pose of multiple people in an image and produces the result correctly. The demerits of this paper is that if there is any rare pose or if the appearance of an image is not good it does not provide the correct output. Also if there is any missing part it provides false detection i.e; it doesn't provide correct key point prediction of the body .So, these demerits are overcome by our work as we are using move net model it helps in providing fast and accurate results.

For multi-person pose estimation, we used a technique called Deep Cut: Joint Subset Partition and Labeling. These researchers are: Eldar Insafutdinov; Siyu Tang; Bjoern Andres; Bjoern Andres; Mykhaylo Andriluka; and Peter Gehler. Partitioning and labeling a collection of body-part hypotheses produced by CNN-based part detectors is the approach adopted here. An

ISSN 1533-9211

DOI 10.5281/zenodo.7127542

integer linear programme formulation of joint detection and pose estimation is a virtue of this study. Coming to the demerits overall Accuracy and Performance are less in 3D pose tracking of multi-person videos. As the accuracy is less we use a 3D model to gain the accuracy in the present work in tracking multiple pose of persons in videos.

Joint multi-person posture estimation and tracking is the next resource to consult. Umar Iqbal, Anton Milan, and Juergen Gall are the books' writers. In multi-person posture estimation and tracking, a Spatio-temporal network may be built together and optimized using integer linear programming. Pose estimation and tracking may be accomplished using this method even when the view is severely obstructed or truncated. These findings are noteworthy for their ability to address a hard problem: joint pose estimation and tracking of an undetermined number of individuals in unconstrained films; there may be any number of persons in a frame of a video for example a frame may containfive members where as another frame may contain ten to twelve members or another frame may contain more than that in this work it is easy to identify the pose of a person although there are uncountable persons in a frame. The demerits are that it is used for only non-commercial purpose, real-time performance may be difficult to achieve on CPU. So, in the present work we will be adding GPU also which helps in making the memory error free and also giving it to access for a limited spacewhich avoids that error.

Implementation

Pose detection is one of the open-source real-time pose detection library which helps in detecting person poses in particular given images or videos. It's a tensor flow's-based pose estimator architecture that can identify joints including the elbows, hips, wrists, knees, and ankles for a single posture or a series of poses. The Pose detection package includes models like Move Net, which operate well on lightweight platforms like browsers or mobile phones. Detecting 17 points and running more than 50 frames per second are the main features of this model. The Lightning and Thunder models are the two available variations of the model.

From the above images the first two images are those points found with a traditional detector and the last two images are that with a move net detector which provides better result than the

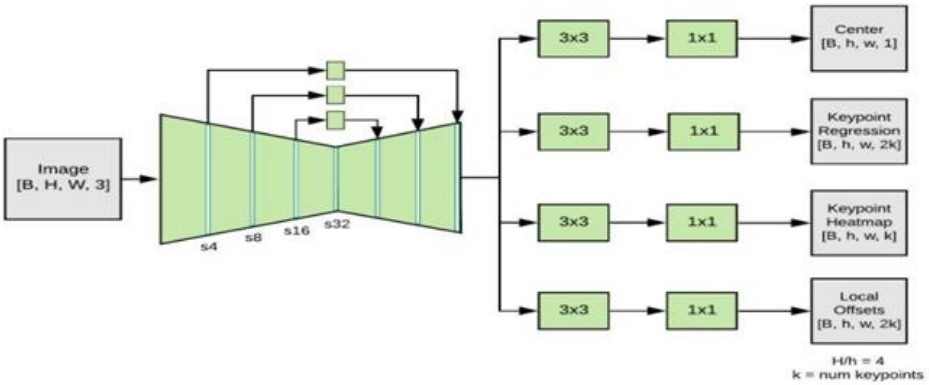
ISSN 1533-9211

DOI 10.5281/zenodo.7127542

original detector. Lightning is mainly made for critical applications. Thunder is mainly used for applications requiring high accuracy.

Move Net is a model used for pose-detection very quickly. It is common for this model to be trained in various fitness, dance, and yoga stances. A technique that uses heat maps to identify critical spots in the human body has been developed. The architecture of this model consists of two components: Retrieval of features it's a MobileNetV2 that's also connected to the internet. It has a high resolution, which may be utilized to get the best possible results. This is a collection of skulls that can foretell the future: Person centre heat map, Key point regression field, Person key point temperature and 2D per-key offset field are all predicted by four prediction heads linked to feature extractor.

Fig: Move net Architecture



The below are the steps used to compute the predictions in parallel:

Step-1: The person's heat map process is accustomed to get the centers of all individual person in the obtained frame. Highest Score location is selected initially. Weight the object center from the person center heat map and compute the value of the heat map.

Step-2: Key point regression output which gives all the set of key points of a person is used for center - out-prediction, which means slice out the key point regression vector at the peak center location.

Step-3: Key points contains pixels and these pixels are multiple by a weight science weight is inversely proportional to the distance. So we are not taking the background people key points. This reduces the scores of the background people key points which concentrates on the single person.

Step-4: We need to find the location of heat map value and add the local 2D offset at that location which is used by the subsequent pixels. The above figure illustrates all the four points.

ISSN 1533-9211

DOI 10.5281/zenodo.7127542

DATASETS:

MPII

The full form of MPII is Max Planck Institute Informatik. This dataset represents Mykhaylo, Leonid, Peter and Bernt which consists of MPII human pose dataset. It consists of 2D pose estimation which consists of 25K videos in which 40K people have body joints annotated on it. It nearly consists of 410 actiities of human which are being performed. Most of the videos are taken from YouTube video with unannotated frames. It even helps in providing the better results in 3D pose test also.

RESULTS

Input A video or an image is given as input to the model which is represented as a tensor of int32 of three dimensions i.e 1x H x w x3 where H and W are the height and width of an image and these are if needed resize to be a maximum size of 256, since the channel order of RGB is [0,255].

Output

The float32 tensor of shape [1, 6, and 56] is the output tensor from the Move Net model. The tensor is illustrated as follows:

1) The first dimension initially is batch size which is always set to value 1.
2) The second dimension is the maximum number of instance detections in the video. The Move Net can detect up to 6 people in the image.
3) The third dimension is the predicted bounding box key points and scores. The key points on the body used are nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle.

The above are the results obtained after using the Move Net model detecting all the key points of a person in any pose. The 17 key points are used and the confident score of each instance is also predicted.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7127542

CONCLUSION AND FUTURE WORK

In this approach we implemented a simple yet effective approach to human body key points tracking invideos. In this we combined the frame-level pose estimation with a quick and effective person level tracking to connect key points over instance of time using Move Net model. The model Move Net has shown better results by providing good accuracy over the whole video. In this we have implemented GPU work which helped in removing the memory error and also helped in setting the memory growth limitedly. In this work the input is given as video and then the video is divided into different numberof frames and the output is obtained by providing key points on the person on frames. So, in the future work we believe that the output can be obtained as video rather than frames with key points and also can use both lightning and thunder models at a similar time to the multi-person domain.

Applications

Human pose estimation is the upcoming and the trending topic in computer vision, it is used in many applications worldwide. Some of them include human-computer interaction, motion analysis, augmented reality and robotics. This is normally used in many different applications in many domain. Some of the most commonly used applications which are used in development fields are

1) Human Activity and movement estimation
2) Some of the human activity applications are sitting gestures detection, cricket umpire signal detection, dance techniques detection etc.
3) Robotics
4) Animation & Gaming
5) Sports

REFERENCES

1) A simple yet effective baseline for 3d human pose estimation ICCV 2017 Julieta Martinez, Rayat Hossain, Javier Romero, James J. Little
2) V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map CVPR 2018 · Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee
3) BodyNet: Volumetric Inference of 3D Human Body Shapes ECCV 2018 · Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, Cordelia Schmid
4) Simple Baselines for Human Pose Estimation and Tracking ECCV 2018 · Bin Xiao, Haiping Wu, Yichen Wei
5) Efficient HR Net: Efficient Scaling for Lightweight High-Resolution Multi-Person Pose Estimation 16 Jul 2020 · Christopher Neff, Aneri Sheth, Steven Furgurson, Hamed Tabkhi
6) Pose Track: A Benchmark for Human Pose Estimation and Tracking Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, Bernt Schiele; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5167-5176
7) Microsoft COCO: Common Objects in Context Tsung-Yi Lin Michael Maire Serge Belongie Lubomir Bourdev Ross Girshick James Hays Pietro Perona Deva Ramanan C. Lawrence Zitnick Piotr Dollar.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7127542

8) 2D Human Pose Estimation: New Benchmark and State of the Art Analysis MykhayloAndriluka1,3, Leonid Pishchulin1, Peter Gehler2, and Bernt Schiele1 1Max Planck Institute for Informatics, Germany 2Max Planck Institute for Intelligent Systems, Germany 3Stanford University, USA.

9) Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affinity fields.

10) In CVPR, 2017 Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images.

11) In ICCV, 2015 J. Donahue, L. A. Hendricks, S. Guadarrama, S. V. M. Rohrbach, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description.

12) In CVPR, 2015 C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect.

13) In ICCV, 2017 G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with R*CNN.

14) In ICCV, 2015 I.Gurobi Optimization. Gurobi optimizer reference manual, 2016.

15) K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes fromvideos in the wild. CRCVTR-12-01, 2012.

16) C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and theimpact of residual connections on learning. In AAAI, 2017.

17) L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B.Schiele. Deepcut: Joint subset partition and labeling for multi-person pose estimation.

18) In CVPR, 2016 D. B. Reid. An algorithm for tracking multiple targets. IEEE Transactions onAutomatic Control, 1979.

19) Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deepconvolutional neural networks.

20) In NIPS, 2012 Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. COCO Dataset. COCO / CC INT'L 4.0

21) Anatomy-aware 3D Human Pose Estimation with Bone-based Pose Decomposition 24 Feb 2020 · Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, Jiebo Luo

22) XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera 1 Jul 2019 · Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, Christian Theobalt

23) Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image ICCV 2019 · Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee.

24) HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation CVPR 2020 · Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, Lei Zhang.

25) "Rotating Solar Trees ", Lecture Notes in Electrical Engineering 601, Springer Nature Singapore Pte Ltd. 2020, and Page No: 482-487.

26) "Late Patterns in Chart Model for Content Examination and Content Mining" IJPT, ISSN: 0975-766X, Volume-8, Issue-2, June-2016, page no: 14729-14736.

27) Text Mining To Knowledge Mining Using Framenet Based Graph Model" IJPT, ISSN: 0975-766X, Volume-8, Issue-2, June-2016, and page no: 14715-14721.