



2D Human Pose Estimation and Activity Recognition Using Machine Learning Techniques

K.Kamaladevi¹,Dr.K.P.SanalKumar²,Dr.S.AnuHNair³Dr. A, Abdul Rasheed⁴

¹ResearchScholar,DepartmentofComputerandInformationSciences,Annamalai University,Chidambaram,India.

²Assistant Professor,

PGDepartmentofComputerScience,R.VGovernmentArtsCollege,Chengalpattu,India.

³Assistant Professor,

DepartmentofCSE,AnnamalaiUniversity,Chidambaram,India[DeputedtoWPT,Chennai].

⁴Director, School of Computing Science, Sree Saraswathy Thiyagaraja College, Pollachi.

E-

Mail:¹nishakanaga@gmail.com,²sanalprabha@yahoo.co.in,³anu_jul@yahoo.co.in,⁴profaar@gmail.com

Abstract:

In order to correctly identify the poses of persons in an image, a technique called "human pose estimate" locates body key points. Human action recognition, sports, tracking, HCI, sign communications, and video surveillance all require this step to be achieved before moving on to the next stage of computer vision. The purpose of this article is to fill in the information vacuum and shed light on the studies of two dimensional human pose estimation. According to the number of persons to be tracked, it can be classified as a single-person or multi-person pose estimation. Afterwards, the many methods for determining a human's position are discussed and several uses and drawbacks are also mentioned. Using a Random Forest model, we address the challenge of 2D human pose estimation in still photos. We propose that picture patches be used to learn a human body's inception module. To extract E-HOG features and train a regression forest, we use patches randomly selected from a bounding box around a specific individual. It's also possible to estimate the joint density function's modes from aggregated leaf samples using an efficient technique. We use three publicly available datasets that include self-occlusion, appearance, and position variants to demonstrate this aspect of our holistic approach. A new dataset is also proposed, which differs from other datasets because of its different resolutions and distortion in the data. A better or similar outcome was reached when we compared our method to current best practices.

Keywords: Human pose estimation, Random Forest, pose estimation and action recognition

DOI Number: 10.14704/nq.2022.20.8.NQ44372

NeuroQuantology 2022; 20(8): 3455-3468

1 Introduction

eISSN1303-5150

www.neuroquantology.com



As depicted in Fig.1, human posture estimation is a hard subject of study in computer vision that seeks to determine the position or spatial location of the body joints of an individuals in a given image or video. A human body with joints and inflexible parts can be estimated by employing image-based observations to determine the pose of an articulated human body [3]. 3D or 2D pose estimate is the process of inferring a person's pose from a photograph. [4]. Approaches to this challenge have been offered in the literature. The visual structures [5–8] were started in the early studies on articulated human pose estimate.

Estimating human body parts from an image is known as human pose estimate, and it involves estimating the location of various elements of a human body from such an image. To put it simply, it's a basic problem for computer vision and has numerous key applications in the fields of sports and action detection as well as character animation and gait analysis in the medical field. Pose estimate is still a challenging topic, despite decades of effort. Human articulation poses a considerable difficulty to pose estimation models.



Figure 1: Deformations among body parts

This is not the case when it comes to selecting the best option from various sources of information. Local deformation costs may be added linearly in linear models, which may lead to a similar deformation score between estimates on the left and right. While it is evident to a human being that the left-hand result is irrational, the right-hand outcome is not. Mixture type and attractiveness score are two other variables that can lead to similar outcomes. Because of this, a non-linear description of distortion, visual score, and mixing type should be constructed.

1.1 Most Popular application of pose estimation are:

- Motion Tracking for Consoles
- Estimation of Human Activity
- Augmented Reality and Motion Transfer
- Elderly Fall Detection
- Robots

1.2 Importance of Pose Estimation

Pose estimation has a number of benefits, including the following: Only bounding boxes can be detected in classical object detection methods (a square). Computers can learn about human body expression by doing pose tracking and detection. In contrast, typical position tracking algorithms are neither quick enough nor resistant enough to partial occlusion to be feasible. Pose monitoring and classification in real-time will be a major influence on future developments in computer vision. Real-time monitoring of a person's position, for example, will help computers better understand human behaviour.

1.3 Motivation of this research work

Human posture assessment has been approached in a variety of ways, depending on how many camera perspectives and input types are used. This job was once restricted to only

one image. A lot of study has been done on how to accurately predict a person's 2D stance by analysing and interpreting their bones. In sports, the most prevalent application of single 2D estimation is posture assessment. In the end, the computer vision domain looked into the matter of estimating many 2D human postures from a single image [5]. Dance, gaming, and yoga are all examples of group activities that can be utilised to help people regain their 2D stance. In the meantime, the topic of 2D human posture estimation has received a lot of interest. The position of the body in 2D space can be retrieved from a single shot using skeleton-based techniques. Indoor applications are the best fit for this technology, which limits the number of sensors that may be used due to interference difficulties. 2D posture estimation in real-world scenarios has received a lot of attention recently. There is a lot of interest in forecasting the 2D posture of sporting event participants, etc.,

1.4 Pose Estimation - Challenges

Pose estimate faces the following difficulties: In human posture estimation, the body's look changes continuously due to various clothing types, arbitrary opacity and background circumstances such as those caused by the viewing angle. Real-world differences like as lighting and weather must be taken into account when determining a person's pose. The identification of perfectly alright joint coordinates is thus difficult for image processing techniques. Following tiny or hardly discernible joint movements might be a challenge.

1.5 Contribution of this research work

This paper has made the following contributions. The nonlinear representation can be constructed from a variety of information sources using a proposed architecture.

- For our initial investigation into human posture estimate, we focus on the 2D space and single human.
- We propose discriminative methods for estimating the 2D body pose of single human from an image using E-HOG features.

- Regressing the body joints of a single individual is done using random forests in order to identify their movements in a frame.

1.5 Research paper organization

In addition, the paper is laid out as follows: Section II includes a brief overview of other books that may be of interest. Further details of the suggested approach are provided here. Detailed explanations of the results and analyses are provided in Section IV. Section V concludes with a summary of the findings.

2 Related Works

Human posture estimation and activity recognition are the two broad categories into which we've organised the approaches shown in this section. There is no room for a lengthy literature review in this short study, so we recommend that readers consult [3, 19] for pose estimation and activity recognition.

Research on 2D human pose estimation using machine learning has just been published [17]. After dividing pose estimate into single and multiple person pipelines, this review generated subcategories for each type. Researchers have now published a new study [18] that examines both 2D and 3D pose estimation using machine learning. There are two categories for 2D human pose estimation: prototype and model-based. In the both cases, methodologies are presented based on these two categories.

Pose Prediction in the Human Body Human posture assessment in the deep learning regime has made tremendous progress in the last five years [30, 28, 6, 31, 19, 11, 32, 20, 22]. Despite the significant performance gains, these prior research focus mainly on enhancing the pose estimate accuracy by applying complicated and extremely intensive models, while largely disregarding the model interpretation cost problem. For real-world applications with limited computational resources, this greatly limits their sustainability and deployability. It is possible to find a few recent studies in the literature that aim to increase model efficiency. For instance, Bulat and Tzimiropoulos developed parametric CNN models for resource-constrained platforms [7]. Due to a

3457



significant performance decrease, this approach is not suitable for long-term use. Accuracy is critical in most circumstances. When it comes to improving efficiency of the model without creating a new algorithm, Rafi et al. [24] used appropriate general-purpose approaches. The exchange among system effectiveness and efficiency cannot be quantified using this approach. These previous techniques, on the other hand, systematically investigate the pose estimation effectiveness issue under the premise of retaining the model's performance rate, so that the resultant model is more useful and trustworthy in real-world application situations.

Pose estimation [14, 32, 33] classify this as a type of holistic recognition. Since there are so many levels of variation in body part articulation, several recent works have employed local body parts [10, 8, 12, 31, 1, 31, 20]. Figure 1 shows how certain techniques have clustered parts into combination of different types since the original work in [10]. The part template can also be warped by varying the sizes and orientations of the warps [12, 20]. In our deep model, we may use the look score, spin, size, and position of these techniques as many sources of information to estimate a person's stance. Models such as tree [1, 36], multi-tree [9], or loopy [9] are used in existing pose estimation techniques for arranging the couple part distortion relationships. Tree modeling are accurate and efficient, but they lack the ability to depict the relationships between body parts in a comprehensive way. The locations of two legs, for example, are autonomous and typically respond to the same visual cue when given the location of a torso in classification trees. A loopy model allows for more complicated interactions between pieces, but it requires imprecise inference. Because of the efficiency of our deep network in both training and testing, we can model even the most intricate interrelationships among our components.

Since G. Hinton's accomplishment in deep learning in [17, 18], deep learning has become

incredibly common. Machine learning algorithms like Support Vector Machine and Boosting are deep models, but they may demand much more computing elements, possibly tenfold more (regardless of input size), than deep models that complexity is matched to the task, according to Bengio [2]. When it comes to classification errors [24], the invariance of source modifications, or modelling multi-modal data [34], deep structure is revealed to be superior. [19, 25, 35, 22, 29, 31, 30, 28, 27] Deep learning has made tremendous advances in computer vision. [3] summarizes recent advancements in deep learning.

To the best of our knowledge, no research has yet been done on developing a machine learning model for estimating human pose. Algorithms that learned from many dimensions such as audio and visual input [34, 16] have influenced our study. Our research focuses on different targeted knowledge from a single image, which is picture data for pose estimation.

3 Methodology

The Random forest model for estimating human position from depth data, random forest has gained considerable traction [6, 12, 17]. In order to extract the human stance from image data, we use a regression forest in this study. We'll go through the fundamentals of a regression forest and how we used them to solve our problem in the sections that follow.

3.1 Regression Forest

For continuous output, the output is estimated using an ensemble of regression trees. An image patch-to-parameter space mapping is the goal of a regression forest training process. Using E-HOG features, the joints define the body's skeletal structure and the picture patches provide an estimate. When it comes to object identification [57], tracking [12], and classification [34], we employ the E-HOG features as descriptions because of their resilience. During training, each tree receives input in the form of a random set of picture patches P and their corresponding skeleton



joint offsets. There are no predetermined locations where the patches are taken from.

The bounding box coordinate system can also be used to express the body pose.

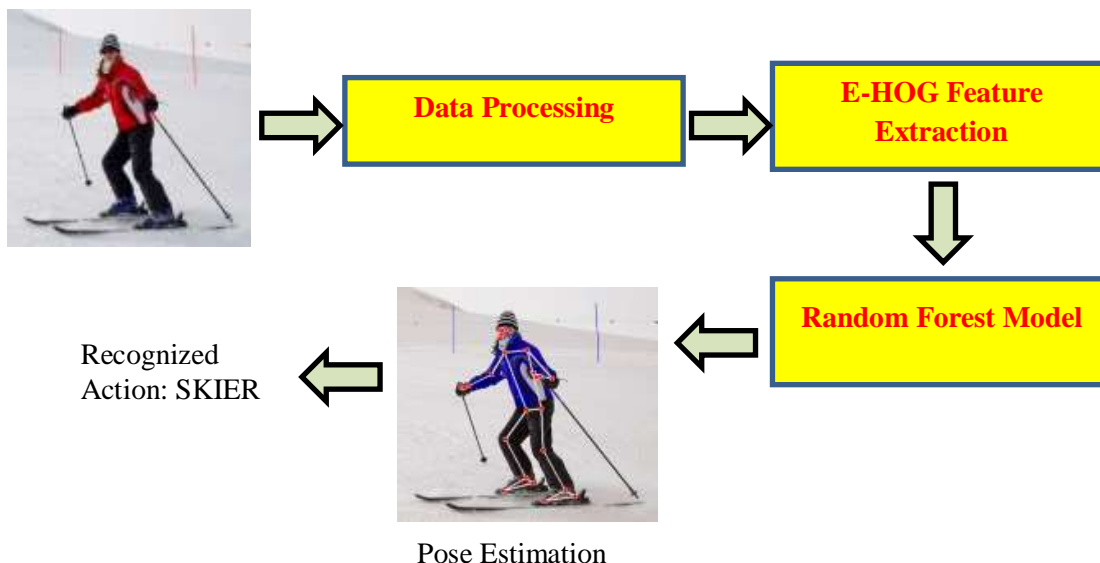


Figure 2: Proposed Architecture

Nodes with binary split algorithms are used to build a tree. q is a split function based on the E-HOG properties of the patch, and is defined for each node. Extraction of the picture patch's E-HOG feature space is done as described in [7].

$$\theta^* = \arg \max_{\theta} g(\theta)$$

where thus, $g(q)$ is the gain in knowledge. The information gain assesses how successfully the split function separates the learning data into two subsets, P_L and P_R . Accordingly, a split

$$g(\theta) = H(p) - \sum_{i \in \{L,R\}} \frac{|p_i(\theta)|}{|p|} H(p_i(\theta))$$

3.2 Method Parameters

A lot of parameters must be established during learning in order to efficiently train the regression forest. We've found that these factors have a significant impact on performance. These are the parameters that we'll be discussing in this section. Patches for Images: During classification and test, the width of all image patches is predetermined. Consequently, the E-HOG features are all of equal dimension. For this, we divide the image gradients into nine bins and follow [7]. Training images appear to be of varying sizes, yet they

In the binary split algorithm, the left P_L and right P_R subsets of a p sample enhanced image are determined. E-HOG's feature vector has a number of dimensions, and the one that yields the best split specifies the split function:

function's selection criteria is to maximise $g(q)$ by splitting input picture patches optimally at the present node. In this formula, the data augmentation is outlined as follows:

are all contained within a same bounding box. Multi scale regression forest learning is not a simple task, according to our findings. Since the bounding box height usually correlates to an individual's height, we grade all data in this way. A uniform scale helps us to represent different human position variations. In the prediction phase, we scale up as well because we presume a local person. A split function is more localized than global, according to our theory of Threshold r . A threshold is used to penalise samples with significant offsets because of this We conducted an experiment in which the

radius was set to 0.8 times the height of the person bounding box, and joints outside of this radius were not included.

3.3 Prediction

A bounding box is used to locate and dynamically resize the individual in the prediction step. Random pixel locations are used as inputs to the regression forest, same like in training. For each random position, a random picture patch is extracted and the E-HOG vector is computed. To get to a leaf in each tree, we use split functions to direct the input picture patch either left or right until it reaches the leaf where we have placed the body joint positions. It is necessary to tally up the results from each tree's leaves after performing this step for a number of random patches. In order to estimate the density function's mode, a collection of aggregated joint offsets must be used to determine which joint offset is the most likely. Mean Shift is the most commonly used algorithm for determining the mode [34]. Given a large number of samples at the leaves, the convergence of Mean Shift is computationally intensive and can take a long time. Using a template matching search, the dense-window technique converges deterministically. Sliding window step and amount of samples are the only variables that affect it.

As a result of the dense-window approach, each grid cell contains the total number of 2D predictions for each joint. This allows for quick estimation. For each joint forecast, the runtime is proportional in terms of s . The votes are then added together in an additive matrix for each individual cell. Together, all of the cells make up a complete picture. Slide the window over the fundamental image to find the window with the most points.

4 Experimental Analysis

Experiments on three tough datasets are presented in this part to demonstrate the effectiveness of our strategy. We demonstrate the adaptability and robustness of our multitasking technique. Pose estimation and activity recognition are the two main difficulties in each of the three subcategories. In both cases, we gauge the effectiveness of our strategy using 2D examples.

4.1 Datasets

Three separate datasets are used to test our method: Datasets like LSP are commonly utilised for human pose estimation because of their accuracy. 1000 photos were used for training and another 1000 were used for testing in the original LSP dataset. Using a person-centric perspective, each image is tagged with 14 joint locations, with the left and right joints uniformly designated.

3460





3461

Figure 3: Sample Images from LSP dataset

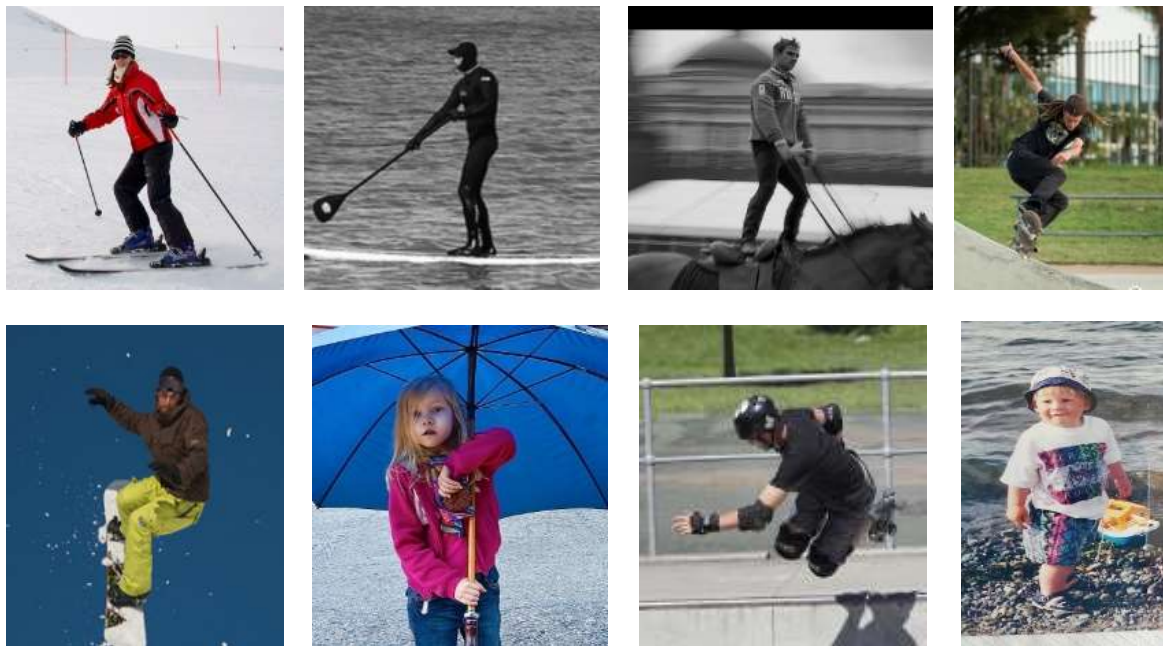


Figure 4: Sample Images from COCO dataset





Figure 5: Sample Images from KTH Football dataset

A huge object recognition, classification, key-point recognition, and captioned dataset, MS COCO (Microsoft Common Objects in Context) There are 328K photos in the dataset. Release of the initial MS COCO dataset took place in 2014. There are 164K photos in total, with 83K in the training set, 41K in the validation set, and 41K in the test set. All the previous test photos as well as 40K fresh images were released in 2015 as part of an extra 81K-image test set. Sample images from the COCO dataset are shown in Figure 4. Football Dataset I from KTH – Multiview. To aid in multi-view restoration, we have annotated a database of football players' joints. Annotated photos of 14 bodily joints are included in the dataset, which comprises 771 photographs of football players. Pictures from the LSP dataset are depicted in Figure 5.

Part-based models have been used for the majority of current methods for estimating human poses from still photos. As a result of our research, we can say with confidence that a more comprehensive approach to human pose estimation yields better results. Here we dissect our comprehensive model, test it against three

different sets of data, and see how it stacks up against other recent techniques. This section begins with an explanation of how we went about computing the regression forest parameters. To avoid parameter overfitting, we run all of our tests only on photos from the LSP dataset's training set. On the KTH Football dataset, we then compare our method against a similar strategy that uses body part categorization through forests [5]. Our model is evaluated on the LSP dataset in order to demonstrate its superiority over part-based techniques. On top of all of that, we've got the brand-new and tough KTH Football dataset. We compare our approach to the part-based strategy to see how it stacks up.

4.2 Evaluation metrics

PCP (% of properly estimated parts) is the standard statistic for human posture estimation in all of our investigations [60]. According to the research, there are two types of PCP scores. It is assumed that the distance between two estimated joint sites and two genuine limb joint positions is at most 50% of a ground truth limb's length in rigorous PCP scoring, while this

3462

distance is only taken into account for limbs defined by two joints in loose PCP scoring. In order to stay up with the associated work, we mostly make use of the loose PCP score in this chapter.

4.3 System Parameters

Using the LSP dataset, we first select the regression forest's parameters [8]. The quantity and density of the trees, and also the width of the picture patch window, are the primary considerations in this study. After analysing the data, we have decided to employ 15 trees with a depth of 40. The trees are incredibly deep because of the wide range of human postures and movements. Each dimension of the patch has been adjusted to 30 pixels in size. Mean Shift and dense-window algorithms were used

to evaluate the prediction stage. We found nearly equal results.

5 Result and Discussion

The part-based strategy that depends on classification forests is compared to our method in this experiment. Each pixel in an image is assigned to a certain bodily joint by the forest in this study. After that, a body previous model (i.e. visual structures) significantly improves the final output. As you can see in Table 1, the findings are rather clear. To implement Yang and Ramanan's [6] technique based on visual structures, we utilized available to the public online code. With the classification, we get comparable results for the majority of body parts. We don't smooth out the findings using a body prior model in our formulation..

3463

Table 1: KTH Football dataset: The evaluation with the loose PCP results

	Head	L_Arms	U_Arms	L_Legs	U_Legs	Fully body	Torso
Proposed Method	0.81	0.92	0.57	0.87	0.97	0.84	0.86
Yang & Ramanan	0.74	0.89	0.55	0.85	0.97	0.80	0.84
Kazemi et al.	0.84	0.93	0.59	0.89	0.98	0.87	0.92
Kazemi et al. + Prior	0.88	0.96	0.60	0.92	0.98	0.91	0.95

For human pose prediction from photos, the KTH Football dataset [8] is one of the most commonly used. We employ it in a variety of ways throughout the thesis. It contains photos of people of various appearances and postures.

Table 2 compares our findings with those of a number of other part-based techniques. Even though we employ a lower quantity of training data than other methods, ours yields outcomes that are comparable to those of the others.

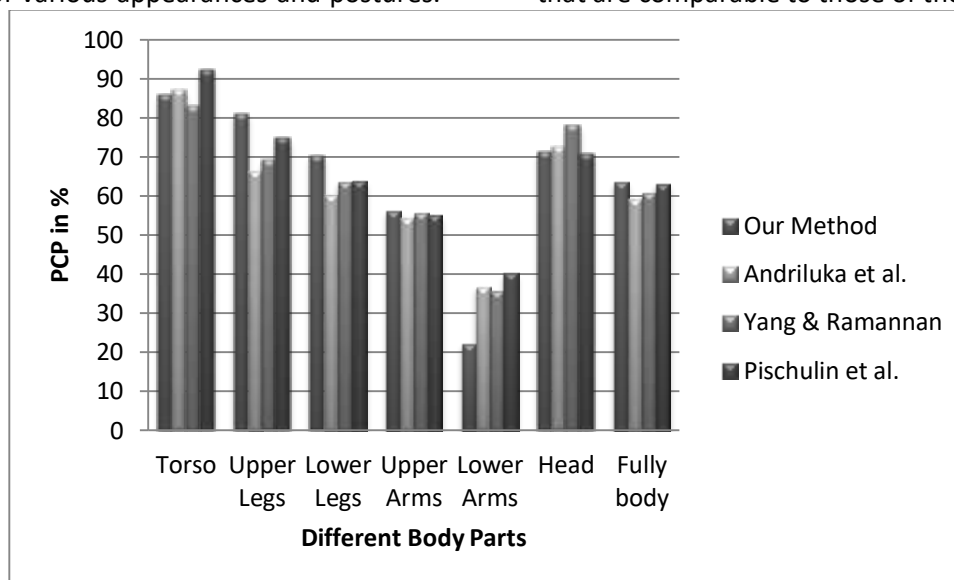


Figure 6: Bar chart for evaluation with the loose PCP results

Regression forests have been trained using a batch of 100 training photos that were simply flipped once to double the training data. In comparison to Pischulin et al., who train with 1000 photos, this is substantially lower. [11] on the other hand, train with 10,000 photos. This is because Random Forest is capable of

generalising to a wide range of positions. Only at the lower arms do we see a decrease in performance due to unclear input. The proposed method keeps track of the competition's results in relation to the other projects.

Table 2: COCO dataset: The evaluation with PCP results

	Head	L_Arms	U_Arms	L_Legs	U_Legs	Fully body	Torso	Head
Proposed Method	71.31	22.14	56.14	70.12	70.12	80.91	63.51	85.81
Andriluka et al.	72.48	36.57	54.39	60.29	60.29	66.31	59.12	86.94
Yang & Ramannan	77.98	35.67	55.47	63.47	63.47	69.11	60.70	82.89
Pischulin et al.	70.81	40.1	54.97	63.74	63.74	74.63	63.10	92.25
Johnson & Everingham	76.91	45.81	67.31	67.11	67.11	74.72	67.42	87.63

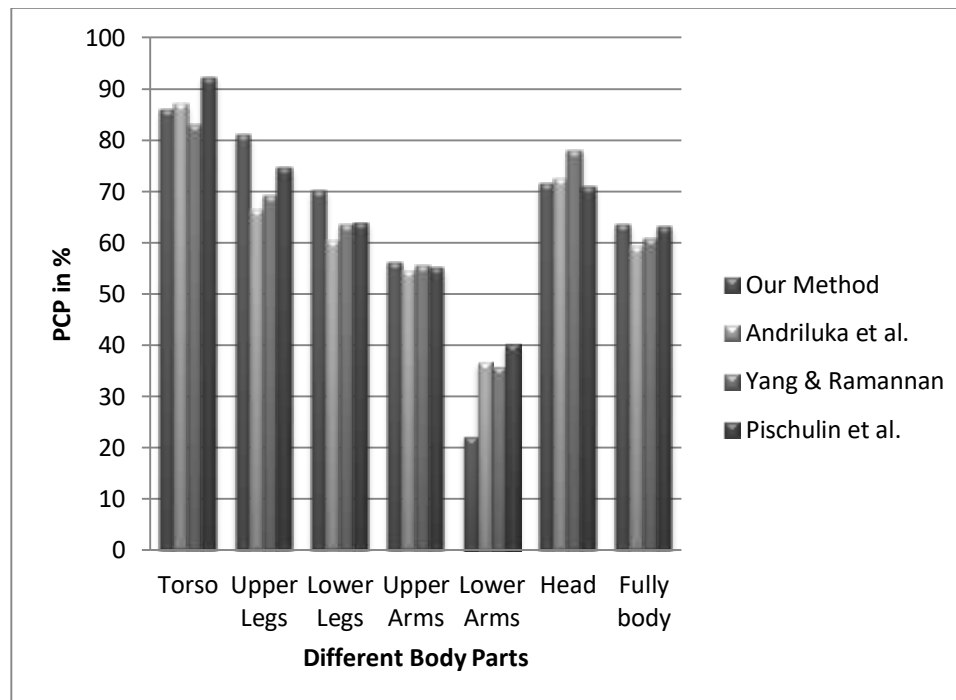


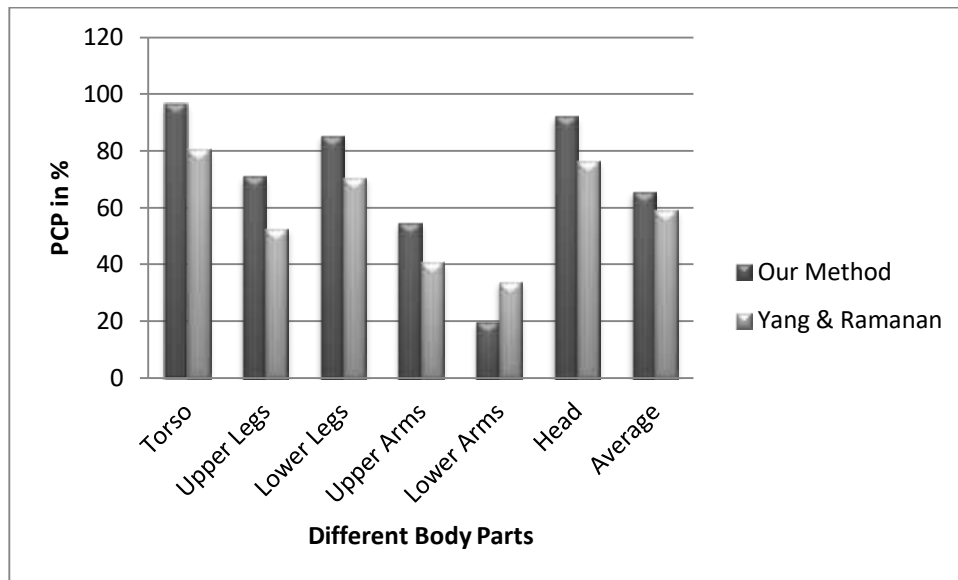
Figure 7: Bar chart for evaluation with PCP results

The KTH Football dataset 1 is a 2D human pose dataset that we recommend for use. The dataset consists of 771 training images of men playing football and 205 testing images of women. The images in this collection are of poor quality and contain a great deal of noise. We show a few examples of the KTH Football dataset and the resulting body positions. We may demonstrate the robustness of our holistic

model by comparing it to input data that contains errors or omissions. The PCP assessment score was used to test our approach on the KTH Football dataset. Table 3 summarises the results, which show that we do better in most body areas but worse in the lower arms. The random forest predicts an average stance when the lower arms are entirely occluded, which is common.

Table 3: LSP: The Evaluation with the loose PCP results

	Head	L_Arms	U_Arms	L_Legs	U_Legs	Torso	Average
Proposed Method	92.11	19.65	54.44	85.12	70.88	96.54	65.44
Yang & Ramanan	76.42	33.75	40.7	70.57	52.43	80.51	59.00



3465

Figure 8: Bar chart for Evaluation with the loose PCP results

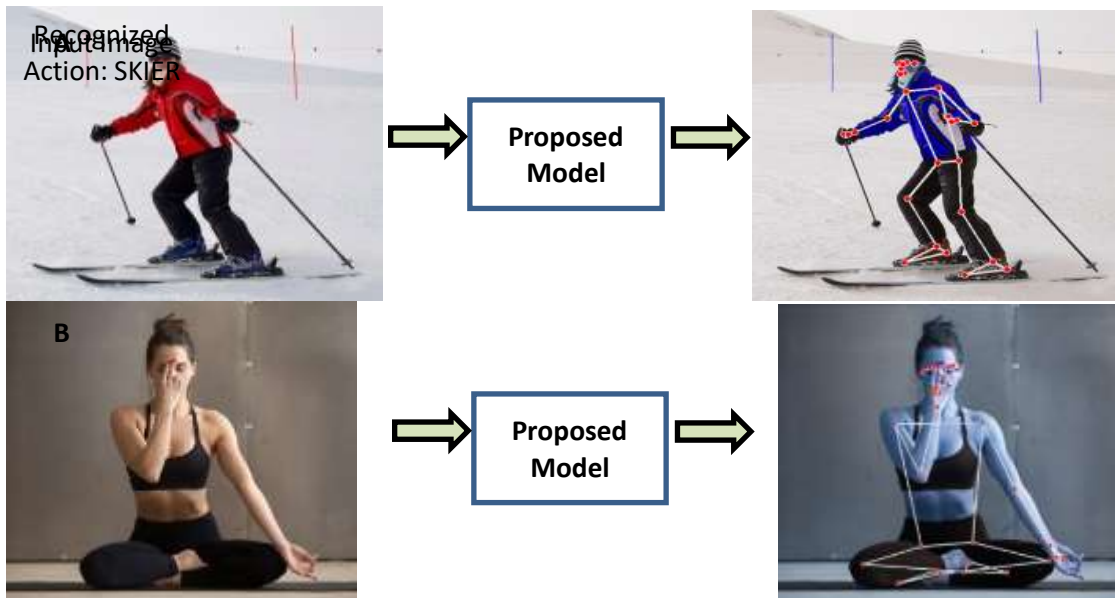


Figure 9: A and B are Pose and activity recognition output from COCO dataset

6 Conclusion

Human pose estimation from 2D photos was provided in this research paper. The model is

based on random forests and E-HOG characteristics extracted from images. On two standard datasets and in comparison to

previous approaches, our model has proven to give good outcomes. We've also included a dataset that's particularly difficult to work with because of the high levels of noise and poor image quality. When compared to the most recent part-based techniques, our holistic approach performs similarly well. The objective is to simultaneously learn a regression model and the characteristics. We use raw data instead of designed features to train features for the problem of human posture estimation. A single-view human posture estimation is still being worked on, however the challenge of 2D human pose estimation is being tackled utilising deep learning. Using Convolutional Neural Networks, a promising deep learning technology, researchers have shown that training a classifier and learning features at the same time can lead to excellent results.

References

1. S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11969–11978.
2. S. C. Babu. (2019). A 2019 Guide to Human Pose Estimation With Deep Learning. [Online]. Available: <https://nanonets.com/blog/humanpose-estimation-2d-guide/>
3. X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in Proc. NIPS, 2014, pp. 1736–1744.
4. D. Mwit. (2019). A 2019 Guide to Human Pose Estimation. [Online]. Available: <https://heartbeat.fritz.ai/a-2019-guide-to-human-poseestimation-c10b79b64b73>
5. M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 1014–1021, doi: 10.1109/CVPR.2009.5206754.
6. M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 623–630, doi: 10.1109/CVPR.2010.5540156.
7. S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in Proc. Brit. Mach. Vis. Conf., 2010, p. 5.
8. L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 588–595.
9. Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in Proc. CVPR, Jun. 2011, pp. 1385–1392, doi: 10.1109/CVPR.2011.5995741.
10. Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2878–2890, Dec. 2013, doi: 10.1109/TPAMI.2012.261.
11. F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 596–603, doi: 10.1109/CVPR.2013.83.
12. M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 723–730, doi: 10.1109/ICCV.2011.6126309.
13. M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (Almost) unconstrained still images," Int. J. Comput. Vis., vol. 99, no. 2, pp. 190–214, Sep. 2012, doi: 10.1007/s11263-012-0524-9.
14. W. Gong, X. Zhang, J. González, A. Sobral, T. Bouwmans, C. Tu, and E.-H. Zahzah, "Human pose estimation from monocular images: A comprehensive survey," Sensors, vol. 16, no. 12, p. 1966, Nov. 2016.
15. H.-B. Zhang, Q. Lei, B.-N. Zhong, J.-X. Du, and J. Peng, "A survey on human pose estimation," Intell. Autom. Soft Comput.,

3466



- vol. 22, no. 3, pp. 483–489, Jul. 2016, doi: 10.1080/10798587.2015.1095419.
16. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep learning for visual understanding: A review,” *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016, doi: 10.1016/j.neucom.2015.09.116.
 17. Q. Dang, J. Yin, B. Wang, and W. Zheng, “Deep learning based 2D human pose estimation: A survey,” *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, Dec. 2019, doi: 10.26599/TST.2018.9010100.
 18. Y. Chen, Y. Tian, and M. He, “Monocular human pose estimation: A survey of deep learning-based methods,” *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897, doi: 10.1016/j.cviu.2019.102897.
 19. A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, “Visionbased hand pose estimation: A review,” *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 52–73, Oct. 2007, doi: 10.1016/j.cviu.2006.10.012.
 20. E. Murphy-Chutorian and M. M. Trivedi, “Head pose estimation in computer vision: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009, doi: 10.1109/TPAMI.2008.106.
 21. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” in *Proc. ECCV, 2014*, pp. 740–755.
 22. M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “MPII human pose dataset,” in *Proc. CVPR, Jun. 2014*, pp. 3686–3693. Accessed: Apr. 13, 2018. [Online]. Available: <http://human-pose.mpi-inf.mpg.de/>, doi: 10.1109/CVPR.2014.471.
 23. M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.
 24. B. X. Nie, C. Xiong, and S.-C. Zhu, “Joint action recognition and pose estimation from video,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1293–1301.
 25. M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, “Learning to refine human pose estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 205–214.
 26. Z. Su, M. Ye, G. Zhang, L. Dai, and J. Sheng, “Cascade feature aggregation for human pose estimation,” in *Proc. CVPR, 2019*, pp. 1–18.
 27. A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Proc. ECCV, 2016*, pp. 483–499.
 28. H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, “Human pose estimation with spatial contextual information,” in *Proc. CVPR, 2019*, pp. 1–10.
 29. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
 30. F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, “Distribution-aware coordinate representation for human pose estimation,” in *Proc. CVPR, 2019*, pp. 7093–7102.
 31. B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in *Computer Vision—ECCV, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018*, pp. 472–487.
 32. Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112, doi: 10.1109/CVPR.2018.00742.
 33. E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, “DeeperCut: A deeper, stronger, and faster multi-person



- pose estimation model,” in Proc. ECCV, 2016, pp. 34–50.
34. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1302–1310.
35. K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5686–5696.
36. K. Simonyan and A. Zisserman, “VGG: Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, pp. 1–14, Apr. 2015.
37. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “AlexNet: ImageNet classification with deep convolutional neural networks,” in Proc. NIPS, 2012, pp. 1097–1105.
38. K. He, X. Zhang, S. Ren, and J. Sun, “ResNet: Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
39. K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988.