A SURVEY ON HUMAN POSE ESTIMATION AND CLASSIFICATION

Suvarna Nandyal[1], Somashekhar S. Dhanyal[*]

[1]Professor, Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalaburgi, Karnataka, India,

[*]Research Scholar, Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalaburgi, Karnataka, India,

---***---

ABSTRACT: *The recent development in estimation and classification techniques in the field human activity recognition has been important progress and outstanding breakthrough. In computer vision, estimating and classifying of human poses is a crucial and difficult task.This paper extensively reviews the recent techniques with different methods and summarizes the benchmark datasets, different evaluation matrices for estimation and classification of human pose. The reviewed technique helps the many applicants to design automated action recognition, activity recognition system and also provides a support for one of the major application, yoga pose recognition system and discuss some of promising future research directions.*

Keywords: Human pose estimation, Action recognition, Activity recognition, Keypoint detection, yoga pose recognition.

1. INTRODUCTION

Human Pose Estimation has aroused computer vision community's interest. It is a challenging step towards understanding object in images/ videos. For decades, the estimation task of human pose has existed, and it aims to determine the position of the human body from given inputs. In recent years, vision-based algorithms have attained rapid progress in a diversity of computer version problems, including picture classification [1], object detection [2], and semantic segmentation [3].Well-designed networks with high estimate competence, better datasets [4, 5], and more real-world investigation of body models [6, 7] have made the estimation tool a lot easier. Action/activity recognition [8], action detection [9], yoga recognition [10], human tracking [11], films and animations, human–computer interface, virtual reality, video monitoring, medical assistance, self-driving vehicles, and sports [12, 13, 14, 15], and so on are all applications of human pose estimation.

There are systems that can distinguish activities from images, however certain activities cannot be detected using only a solitary stationary image from the present. There is a requirement for additional information pertaining to the events that occurred prior to and following the occurrence. Videos are more accurate in recognizing actions from videos in this situation.

Apart from many applications the Human Pose approximation is needed and necessary in the arena of yoga recognition also because an alternative and complementary medicine for its beneficial effects on physical and mental health. Numerous studies showed that the properties of Yogasan improves health from different disorders and reduce mental stress. There are many ways to perform to practice yoga, by attending the classes manually or online sessions or by self-learning through reading books, watching videos through internet. Due to many reasons people are attracted towards the self-learning. Although self-learning is crucial in yoga practice, incorrect postures can cause major injury to the body's muscles and ligaments [44].Many practitioners believe that the development of computer-assisted training methods will help them improve their performance while also protecting them from damage [16], thus requires an automated yoga recognition system for self-learning .To develop automated human/yoga recognition system there are also many challenges in the domain. A good recognition technique should be able to distinguish between actions from different classes while also generalizing across variants within a single class. This will become more difficult as the number of action classes grows, as the overlap between classes grows. Thesituation in which the action takes place contributes significantly to variation in the recording; changing lighting conditions and dynamic backgrounds make camera issues even more difficult. The paper intends to present a synoptic review of human pose estimation and classification in order to develop an automated yoga recognition system as the initial step towards the development of self-training system. We outline the human pose Estimation approaches, human body models, methods and describe the datasets of vision based human/yoga pose classification from images/videos (A video is nothing more than a sequence of images), classification methods, some evaluation matrices and addresses the future directions.

2. APPROACHES FOR DETERMINING HUMAN POSE

The estimation of human posture is a crucial stage in recognizing human pose, and it seeks to obtain the human body's posture from given inputs (image/video). Human Pose estimation approaches are divided into two categories. Based on geometrical projection or a specific image processing challenge, generative and discriminative approaches are divided into two categories. Another method to categorize the human pose estimate problem is whether it is approached from a high-level abstraction and worked down or from low-level pixel evidence and worked up. Top-down procedures are those that work downwards, whereas bottom-up approaches work upwards [17].

2.1 Generative: Creative strategies can be managed in unique ways constructed on a variety of demonstrations of human body prototypes, such as preexisting conceptions near the body model's construction, geometric estimate from dissimilar perspectives to 2D or 3D space, and so on. This approach is not considered to be more computationally efficient than the discriminative approach [18] since the generative exemplary is given in terms of a computer graphics representation of postures.

2.2 Discriminative: If you don't want to use human body models, you may either directly train a mapping between input sources and the space of human poses (learning-based)or browse through examples that already exist (example-based).Starting with image evidence, discriminative techniques estimation pose via a mapping or a search-based mechanism [18]. After training the prototypical, test is substantially quickerthanreproductive approaches since it delves into construction calculations or limited search problems instead of maximizing a high-dimensional parametric domain. Within their scope, the method looks for the best solutions [19].

2.3 Top-down: To begin, use a high abstraction level to recognize people and produce their positions in bounding boxes. After that, each individual goes through the problem-solving process of changing from high-level semantics to lower-level video evidence, with high-level semantics guiding low-level identification. The computational cost of top-down techniques increases dramatically as the number of persons in an image grows [20].

2.4 Bottom–up: Predict all of each person's body components in the input image first, then aggregate them using human body model suitable or other techniques. Depending on the procedure, little template patches, joints, limbs may be used. In these techniques, fragments of visual evidence are collected and labeled to provide descriptive features. These traits are at times used to forecast human postures directly, and they are now and then used to detect body components whose manifestations in photos are subsequently combined to generate a human existence [18].The computing cost will remain steady as the number of people in an image grows.

Bottom-up approaches, on the other hand, have difficulty grouping matching body parts if there are certain people with a lot of overlap [21].

3. HUMAN BODY MODELS

Human Pose estimation is a computer vision-based system that recognizes and analyzes human posture. Thedemonstrating of the human body is the most significant aspect of human pose approximation. The human body is a complex and flexible non-rigid object with a variety of distinguishing characteristics, including kinematic structure, body shape, and surface texture, as well as location of body components or joints [1].The three most prevalent types of human body models are shown in Fig.1: skeleton-based, contour-based, and volume-based models.

3.1 Skeleton-based Model: A human body's skeletal system is made up of joints (key points) such ankles, knees, shoulders, elbows, wrists, and limb orientations. This model is used in both 2D and 3D human posture estimation methods [22] due to its adaptability. A skeleton-based model is a collection of joint positions and limb orientations that follow the skeletal structure of the human body (typically between 10 and 30).Also labeled as a graph with vertices signifying joints and edges expressing limitations or prior relationships of joints inside the skeletal structure [23].This simple and flexible human body topology is frequently used in both 2D and 3D human pose estimates and human pose datasets [24].

3.2 Contour-based Model: Widely employed in human posture estimate methods, this image comprises approximate width and constructed by placing for the body's limbs and torso, and depicts human body components as rectangles or the person's silhouette's borders. The model, which also includes cardboard models and Active Shape Models (ASMs) [25], is a useful way to portray a human posture.

3.3Volume-based Model: To depict 3D human body silhouettes and positions, geometric shapes or meshes are employed. Cylinders, conics, and other geometric shapes were used to model bodily parts in the past. [26]. Modern volume-based models, which are often obtained via 3D scans, are represented as meshes. Popular volume-based models include Shape Completion and Animation of People (SCAPE) [27], Skinned Multi-Person Linear Model (SMPL) [28], and a unified deformation model [29].

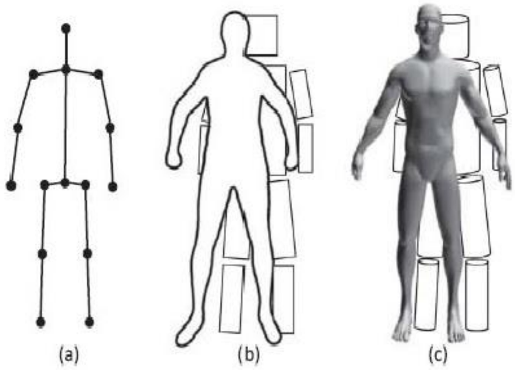


Figure 1.Human Body Models (a) skeleton-Based Model (b) counter-based model (c) volume-based Model

4. KEYPOINT DETECTION METHODS

Key point detection entails detecting people while also locating their important points. The terms "key points" and "interest points" are interchangeable. They're spatial positions, or points in an image, that describe what's fascinating or stand out in the image/video, and they express a graphical representation of a person's orientation. It is essentially a series of coordinates that may be joined to define a person's pose. Each skeletal co-ordinate is referred to component (joint or key point).Pair is a valid connection between two pieces (or a limb). Individual parts are generally identified first, then connections are formed between them to produce the pose in the manner listed below.

4.1 OpenPose: Like many previous bottom-up algorithms for estimation of many human poses, Open Pose begins by detecting parts (key points) corresponding to each person in the image, then allocates parts to distinct human. OpenPose model's architecture is depicted in Fig 2.Using first few layers, OpenPose network collects features from an image (VGG-19 in the above flowchart).After that, the features are routed to two convolution layer branches that run parallel to one another.

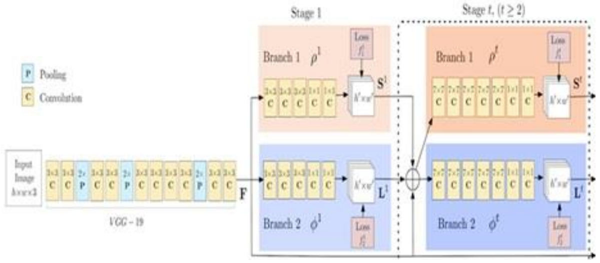


Figure 2: OpenPose architecture

The first branch generates 18 confidence maps, which each represents a different aspect of the human posture skeleton. According to the second branch, a collection of 38 Part Affinity Fields (PAFs) will be generated, which will be used to establish relationships between parts in the model.

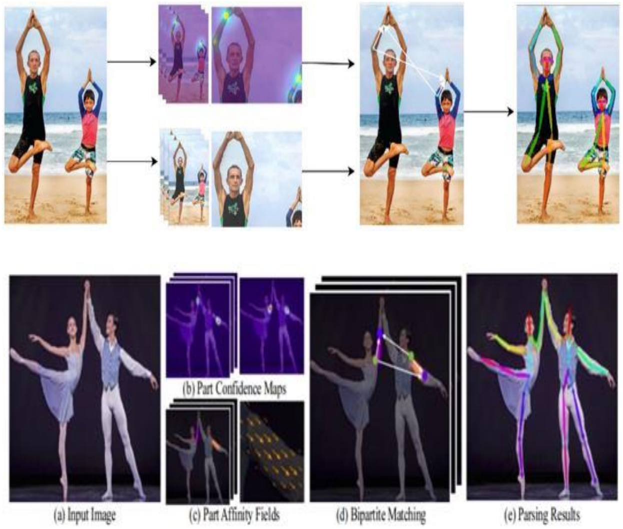


Figure 3.OpenPose steps for estimating human pose.

Each branch's forecasts are refined over successive stages. Part confidence maps are used to create bipartite graphs between pairs of parts, as seen in fig 3.Weaker linkages in bipartite graphs are trimmed using PAF values. Using the processes discussed above, human position skeletons can be approximated and allocated to each person in the image.

4.2 DeepCut: The following challenges were identified using bottom-up method for multi-person human pose approximation:

i. Make list of potential D body part candidates. For each person in the image, this collection represents all conceivable body part locations. Choose a subset of bodily parts from the list of choices above.

ii. One of the C body part classes should be assigned to each body part. The numerous types of body parts are represented by the different body part classes, such as "arm," "leg," "torso," and so on.

iii. Parts of the body that belongs to the same individual but are separated.

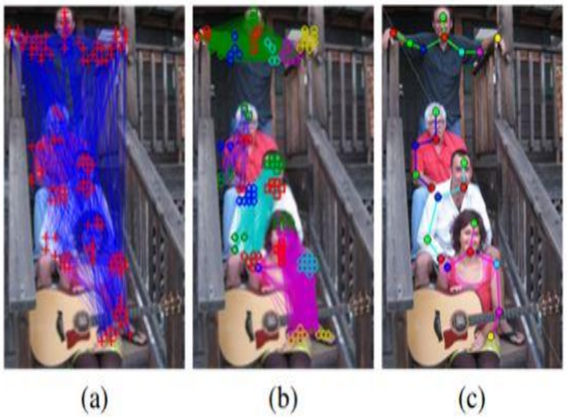


Figure 4: The approach is depicted graphically. (a) Initial detection (b) jointly clustered (c) Predicted pose sticks

4.3 RMPE (AlphaPose): Pose Estimation from the top down is a frequent method. The location of the subject is utilized to calculate their pose. The posture extraction approach may perform badly due to issues with localization and duplicate bounding box forecasts.

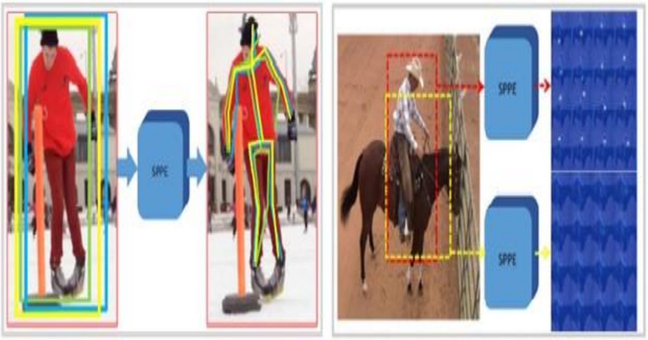


Figure 5: Effect of redundant predictions (left) and bounding boxes with a low level of confidence (right)

The authors recommended that an erroneous bounding box be used to abstract a high-quality solitary person region using the Symmetric Spatial Transformer Network (SSTN).Human posture skeleton for that person in this extracted region is estimated using a Single Person Position Estimator (SPPE).A Spatial De-Transformer Network is utilized to remap the estimated human position to the image's original coordinate system (SDTN).Finally, to solve issue of duplicate pose deductions, a posture that is parametric the technique of Non-Maximum Suppression (NMS) is used, as illustrated in Fig 5.

4.4 Mask RCN: For semantic and instance segmentation, this is a well-known design. The model predicts the image's many elements' bounding box coordinates, as well as a mask that conceptually segments the item. Fig 6 shows how the basic design can be easily modified to estimate human position.

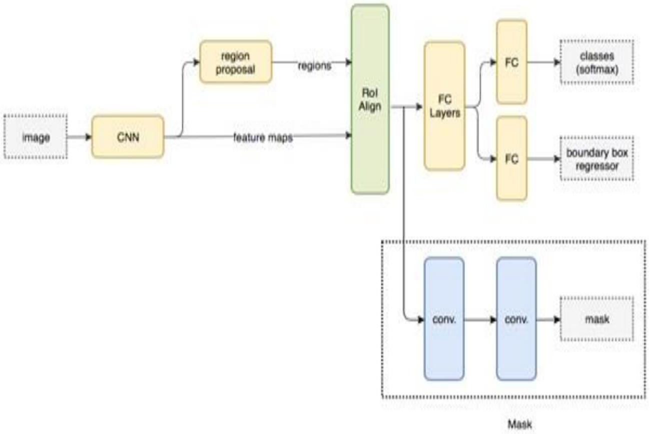


Figure 6: Mask RCNN Architecture

To extract feature maps from an image, the basic design starts with a CNN.A Region Proposal Network (RPN) uses these feature maps to find bounding box contenders aimed at the existence of objects. Candidates for the bounding box select a region (area) from the CNN-extracted feature map as their starting point. RoIAlign is a layer that is used to minimize the size of the removed feature and make it more consistent because the bounding box contenders can be of different sizes. This mined feature is now sent through parallel CNN branches, which predict the bounding boxes and segmentation masks at the end. This strategy is quite similar to the top-down method.

5. DATASETS

As we can see, there are a plethora of publicly available datasets that allow us to compare diverse approaches and gain perception into the (in) capabilities of various approaches. The most often used sets are discussed.

- **COCO [4]:** The COCO key - point database is a multi-person two-dimensional pose estimation dataset based on Flickr. COCO is the world's largest 2D Pose Prediction dataset, which is used to test algorithms for 2D Pose Estimation.

- **MPII[30]:** The MPII humans posture database is a multi-person two-dimensional pose estimation dataset gathered from YouTube videos. It covers approximately 500 different human activities. MPII was the first dataset to have such a huge number of distinct positions.

- **HumanEva [31]:** Human Eva is a dataset for single-person 3-dimensional Pose Estimation that contains video sequences captured with a mixture of RGB and grayscale camera. HumanEva was first large-scale 3D Pose Estimation dataset. Marker-based motion capture (mocap) cameras are used to capture real-world 3D poses.

- **KTH [32]:** The KTH human motion dataset features 25 distinct people performing six different movements (walking, jogging, sprinting, boxing, hand waving, and hand clapping).Outdoors, outdoors with zooming, outdoors with various attire, and interiors are the four scenarios used. There is a lot of diversity in the quality and length of the performance. The backgrounds are quite consistent. There is only minor camera movement aside from the zooming scenario.

- **Weizmann[33]**: The Weizmann Institute's human action dataset contains ten actions done by ten people (walk, run, leap, gallop sideways, bend, one-hand wave, two-hand wave,jump in place, jumping jack, and skip). The backgrounds are static, while the collection includes foreground silhouettes. The point of view is fixed.

- **INRIA XMAS [34]:**Introduced the IXMAS dataset, which covers behaviors from five different perspectives.Around11 people participated in fourteen acts (check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw overhead and throw from bottom up).In terms of camera configuration, the movements are carried out in any direction. The camera perspectives are set in stone, with a motionless background and lighting. There are silhouettes and volumetric voxel renderings in the collection.

- **UCF [35]:** Around 150 sequences of sport motions in the UCF sports activity dataset. Bounding boxes for human figures are part of the collection.For the majority of action classes, there is considerable variation in action execution, human appearance, camera movement, perspective, illumination, and background.

- **YOGA-82 [36]**:Existing pose estimate dataset's limits in terms of pose variety have prompted researchers to dig deeper and offer a novel idea of fine-grained hierarchical pose categorization, together with a diverse dataset of 82 complicated poses.

- **YogaVidCollected [10]**YogaVidCollected is a data set of yoga videos that can be used to recognize yoga poses. There are 6 asana in all, 88 videos, and 1.1 GB of data. At 30 frames per second, the total time is 1 hour, 6 minutes, and 5 seconds.

6. HUMAN POSE CLASSIFICATION METHODS

Classification is the domain plays the most important role of image analysis. Classification accepts the given input images/videos and produces output in the form of classifying the human or yoga pose. There are many classification methods, some are listed below.

6.1 Support Vector Machine (SVM): Data analyzed using a machine learning technique for categorization and regression analysis. A supervised learning method is SVM. The categorization is accomplished by constructing a sequence of hyper planes in a multidimensional space with a big margin between the elements of various classes. It does quite well in terms of recognition and classification. As a result, it is by far the most popular classifier.

6.2 Naive Bayesian classifier: Based on Bayes' theorem [37], a supervised learning algorithm and probabilistic classifier. It is a technique that involves counting the instances in which a particular motion occurs in a video system. To distinguish a novel action, the Bayes rule is applied and activity with highest a posteriori probability is picked.

6.3 K-Nearest Neighbor: The supervised learning domain includes one of the most basic yet fundamental classifications. It entails classifying objects based on votes the majority of their neighbors [38].Features are represented as location vector in a multi - dimensional space of features to locate neighbors.

6.4 K means: It is data structure-sensitive clustering technique that iteratively calculates the k-distances from each class centroid to each datum [39].Unsupervised classification algorithm K-means divides items into k groups depending on their attributes. The sum of the distances between each item and group or cluster centroid is minimized during grouping. It is most suited for classifying human positions.

6.5 Mean shift clustering: Image processing and computer vision are two fields where it can be used. The algorithm continually assigns each piece of data to the nearest cluster centroid given a series of data points, with direction to a nearest cluster centroid determined by the location of the majority of nearby points.

6.6 Machines finite state: Static motions and postures are represented by states, transitions are used to provide temporal and/or probabilistic limitations, and arcs between the beginning and final states are used to describe dynamic gestures.

6.7 Hidden Markov Models: This model addresses the keys to the segmentation problem. They're a type of Markov chain [40], which are determinate state automata with probabilities on each arc.

6.8 Dynamic time warping: Based on their related characteristic values, this technique determines the distances between every pair of potential locations in two signals. It is utilized to overcome problems of rate variations caused by the identification of human poses by calculating and recognizing motion in a video stream [41].

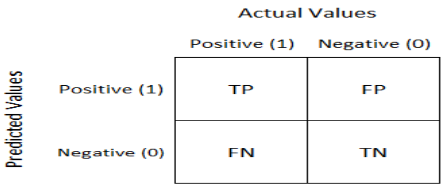
6.9 Neural networks: Mathematical models whose strategy is based on biological neuron functioning [42].Probabilistic learning approaches, particularly Bayesian, are commonly used to optimize them. Neural networks make it possible to create quick classifications that may be used in real-time systems. Neural networks are also the best choice for recognizing poses.

7. EVALUATION METRICS

For human posture recognition, numerous performance criteria used in various classification disciplines have been modified and applied. We cite frequently used measures like accuracy, precision, and recall in this section based on [43].Before summarizing these metrics, let's define True Positive, True Negative, False Positive, and False Negative in the framework of posture recognition (see fig 7):

- **True Positive** actions are those in which the actual and projected transactions coincide.
- **False Negative** actions are those that belong to a specific class but are predicted not to be from that class.
- **True Negative** actions are those in which the actual and projected transactions do not match the search class.
- **False Positives** are activities in which the actual transactions don't match the examined class, but are projected to do so.

7.1 Sensitivity: It's also known as recall, true positive rate, or detection probability. It refers to positive cases that were actually predicted to be positive. Sensitivity in posture identification refers to the proportion of poses predicted in each class.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 7: Structure of the confusion matrix

Similarly, (1 - sensitivity) determines the system's inability to detect pose.
This can be stated mathematically as:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

7.2 Precision: It's also identified as the Positive Prediction Value (PPV), and it refers to probability of a detected example of pose occurring in real life. The likelihood of the recognizer erroneously identifying a detected posture is also determined by (1 - accuracy).

This can be stated mathematically as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

7.3 Specificity: It's sometimes referred to as the false positive rate (FPR) or the true negative rate (TNR). It's based on actual negative cases that were anticipated to be negative. It assesses the sensitivity of the system to negative classes. This can be stated mathematically as:

$$Specif\ icity = \frac{True\ Negative}{True\ Negative + False\ Positive} \qquad (3)$$

7.4 Negative Predictive Value (NPV): It's known as "negative precision" since it calculates the probability that a negative identification is correct when compared to all previous negative identifications.

This can be stated mathematically as:

$$NVP\ = \frac{True\ Negative}{True\ Negative + False\ Negative} \qquad (4)$$

7.5 F_Measure: It provides details on the test's accuracy. It calculates the precision and recall harmonic mean. As a result, the F measure defines how precise and resilient the classifier is at the same time. Its best value is 1 and its worst value is 0.

This can be stated mathematically as:

$$F\_Measure = 2 * \frac{Precision\ *\ Recall}{Precision +\ Recall} \qquad (5)$$

7.6 Accuracy: When the classes are evenly sampled, the accuracy produces good results. Calculated is the proportion of valid predictions in respect to the total number of samples. This can be stated mathematically as:

$$Accuracy = \frac{NumberOfCorrectPredictions}{TotalNumberOfPredictionsMade} \qquad (6)$$

$$Accuracy = \frac{TrueP\ ositive\ +\ TrueNegative}{TotalNumberOfSamples} \qquad (7)$$

7.7 Likelihood Ratio: Compares the likelihood of an action predicted correctly vs the likelihood of an activity predicted incorrectly. It can be used to calculate both true positive and true negative outcomes.

Mathematically, this can be expressed as:

$$LR{+}{=}\frac{Sensitivity}{1\ -\ Specificity} \qquad (8)$$

$$LR{-}{=}\frac{1 - Sensitivity}{Specificity} \qquad (9)$$

7.8 Confusion Matrix: It is also known as the error matrix, and it summarizes the prediction outcomes. It also describes the model's overall performance. The classifier's errors, as well as their types, are displayed in the confusion matrix. The matrix's rows indicate expected class instances, whereas the columns represent actual class instances, or vice versa (Fig.7).

7.9 Intersection over Union (IoU): Also known as the Jaccard similarity coefficient or the Jaccard index. It assesses the detector's accuracy on a certain dataset. The region of intersection between the predicted bounding box and the ground-truth bounding box is referred to as "Area of overlap," while the area contained by the predicted bounding box and the ground-truth bounding box is referred to as "Area of Union."

$$IoU = \frac{Areaof\ overlap}{Areaof\ union} \qquad (10)$$

8. CONCLUSION

The purpose of this paper is to present the most recent work in this field of study. It initially discuss the objective of human pose estimation and then it presents the human pose estimation approaches and key point detection methods for pose representation and also discussed about some of common datasets, different type of classification methods provides an effective survey study to design and develop the automated human recognition system.In numerous computer vision applications, the necessity to interpret human action has become unavoidable. On the other hand, the major application fields yoga pose recognition becoming a prominent research work because of developed techniques to perform yoga in front of system without the help of trainer and become self-learner. We conclude that our study helps to design and develop the automated human recognition system/Yogasan recognition system with different poses irrespective of many challenges.

REFERENCES

1. Ali S, Shah M. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010,32(2),288-303.DOI:10.1109/TPAMI.2008.284.
2. Anguelov D, Srinivasan P,KollerD,ThrunS,Rodgers J, Davis J. Scape: shape completion and animation of people. *Journal of ACM Transactions on Graphics*. 2005,24(3), 408–416.https://dl.acm.org/doi/10.1145/1073204.1073207.
3. Akansha A, Shailendra M, Singh N. Analytical review on video-based human activity recognition. 3rd International Conference on Computing for Sustainable Global Development. 2016, 3839–3844.

4. Andriluka M, Iqbal U, Milan A, Insafutdinov E, Pishchulin L, Gall J, Schiele B. Posetrack: A benchmark for human pose estimation and tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2018, 5167–5176.https://arxiv.org/abs/1710.10000.
5. Andriluka M, Pishchulin L, Gehler P,Schiele B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2014, 3686–3693.DOI: 10.1109/CVPR.2014.471.
6. Bai L, Efstratiou C,Ang S. WeSport: utilising wrist-band sensing to detect player activities in basketball games.IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops).2016,1-6,DOI: 10.1109/PERCOMW.2016.7457167.
7. Ben M, Trabelsi I,Bouhlel M. Human action analysis for assistance with daily activities. *International Journal on Human Machine Interaction*. 2016,1-10, https://www.researchgate.net/publication/305848684.
8. Liefeng B, Cristian S. Twin Gaussian Processes for Structured Prediction.*International Journal of Computer Vision*. 2010, 87, 28–52.DOI 10.1007/s11263-008-0204-y.
9.Bhardwaj R, Singh K. Analytical review on human activity recognition in video. 6th International Conference Cloud System and Big Data Engineering. 2016, 531–536.https://ieeexplore.ieee.org/document/7724978.
10. Cao Z, Simon T, Wei E, Sheikh Y. Real-time multi-person 2d pose estimation using part affinity fields. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017,1302-1310.DOI: 10.1109/CVPR.2017.143.
11. Chen T, He Z, Hsu C. Computer-assisted yoga training system. *Multimedia Tools and Applications*, 2018,77, 23969–23991.https://doi.org/10.1007/s11042-018-5721-2.
12. Chen T.He Z, Hsu C,ChouL,Lee Y, Lin P. Yoga Posture Recognition for Self-training. In Proceedings of the 20th Anniversary International Conference on Multimedia Modeling 2014,8325, 496–505.https://doi.org/10.1007/978-3-319-04114-8_42
13. Cootes F, Taylor J, Cooper H and Graham J. Active shape models-their training and application. *Journal of Computer Vision and Image Understanding*, 1995,61(1),38-59.https://doi.org/10.1006/cviu.1995.1004.
14. Felzenszwalb F, Huttenlocher P, Pictorial structures for object recognition. *International Journal of Computer Vision*.2005,61,55–79.https://doi.org/10.1023/B:VISI.0000042934.15159.49.
15. Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes .Tenth IEEE International Conference on Computer Vision. 2005,1,1395-1402.doi: 10.1109/ICCV.2005.28.
16. Insafutdinov E, Andriluka M, Pishchulin L, Tang S, Levinkov E, Andres B, Schiele B. ArtTrack: Articulated Multi-Person Tracking in the Wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, 1293-1301, doi: 10.1109/CVPR.2017.142.
17. Joo H, Simon T, Sheikh Y. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In Proceedings of Conference on Computer Vision and Pattern Recognition. 2018, 8320-8329.DOI: 10.1109/CVPR.2018.00868.
18. Joo H, Simon T, Li X, Liu H, Tan L, Gui L, Banerjee S, Godisart T.S, Sheikh Y. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017,190–204. https://doi.org/10.1007/s41095-020-0171-y.
19. Ju X, Black J, YacoobY. Cardboard people: A parameterized model of Articulated image motion. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition. 1996, 38–44. https://doi.org/10.1007/s11263-006-9781-9.
20. Kanazawa A, Black M.J, Jacobs D.W, Malik J. End-to-end recovery of human shape and pose. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2018, 7122–7131. DOI: 10.1109/CVPR.2018.00744.
21. Krizhevsky A, Sutskever I, Hinton G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems. 2012,1,1097–1105. https://doi.org/10.1155/2020/7607612.
22. Liu Z, Zhu J, Bu J, Chen C. A survey of human pose estimation: the body parts parsing based methods. *Journal of Visual Communication and Image Representation*. 2015, 32,10–19.https://doi.org/10.1016/j.jvcir.2015.06.013.
23. Li M, Zhou.Z, Li J, Liu X, Bottom-up pose estimation of multiple person with bounding box constraint, 24th International Conference on Pattern

Recognition. 2018,115-120. DOI: 10.1109/ICPR.2018.8546194

24. Long J,Shelhamer E, Darrell T, Fully convolutional networks for semantic Segmentation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2015,3431–3440. DOI: 10.1109/CVPR.2015.7298965.
25. Loper M, Mahmood N, Romero J, Pons-Moll G, Black M.J. Smpl: A skinned multi-person linear model. ACM Transactions on Graphics.Vol 34.Issue 6.pp.1–16, https://doi.org/10.1145/2816795.2818013.
26. Luvizon D.C, Picard D, Tabia H, 2d/3d pose estimation and action recognition using multitask deep learning. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2018, 5137–5146. https://hal.archives-ouvertes.fr/hal-01815703.
27. Manisha V, Sudhakar K, Yuta N, Shanmuganathan R. Yoga-82: A New Dataset for Fine-grained Classification of Human Poses. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020, 4472-4479, DOI:10.1109/CVPRW50498.2020.00527.
28. Minnen D, Westeyn T, Starner T, Ward J, Lukowicz P, Performance metrics and evaluation issues for continuous activity recognition. Performance Metrics for Intelligent Systems. 2006, 4.303-317, DOI=10.1.1.129.5025.
29. Nguyen-Duc-Thanh N, Stonier D, Lee S, Kim H, A new approach for human-robot interaction using human body language. In Proceedings of the 5th international conference on Convergence and hybrid information technology, 2011, 762–769, DOI: 10.1007/978-3-642-24082-9_92.
30. Nordsborg B, Espinosa G, Thiel V. Estimating energy expenditure during front crawl swimming using accelerometers. Procedia Engineering. 2014,132-137, https://doi.org/10.1016/j.proeng.2014.06.024.
31. Pai F, ChangLiao H, Lin K P, Analyzing basketball games by a support vector machines with decision tree model, *Neural Computing and Applications*, 2017, 28,4159–4167. https://doi.org/10.1007/s00521-016-2321-9.
32. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the in Neural Information Processing Systems. 2015, 28,91–99.DOI: 10.1109/TPAMI.2016.2577031.
33. Rodriguez D, Ahmed J, Shah M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. IEEE Conference on Computer Vision and Pattern Recognition. 2008, 1–8.DOI: 10.1109/CVPR.2008.4587727.
34. Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local svm approach. In Proceedings of the 17th International Conference on Pattern Recognition, 2004,3.32–36. DOI: 10.1109/ICPR.2004.1334462.
35. Shan Z, Su E, Ming L. Investigation of upper limb movement during badminton smash. In Proceedings of 10th Asian Control Conference (ASCC). 2015,1–6, DOI: 10.1109/ASCC.2015.7244605.
36. Sidenbladh.H, Black.M, Fleet D, Stochastic tracking of 3D human figures using 2D image motion. In Proceedings of 6th European Conf. Computer Vision. 2000, 1843,702-718. https://doi.org/10.1007/3-540-45053-X_45.
37. Sidenbladh H, De la T, Black M. A framework for modeling the appearance of 3d articulated figures. In Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition. 2000, 368-375, DOI: 10.1109/AFGR.2000.840661.
38. Sigal L, Balan O, Black J, Humaneva. Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 2010, 87(1), 4-27.https://doi.org/10.1007/s11263-009-0273-6.
39. Torres F,Kropatsch G.Top-down 3D Tracking and Pose Estimation of a Die Using Check-Points". Structural, Syntactic, and Statistical Pattern Recognition.2012.pp.492-500. DOI: 10.1007/978-3-642-34166-3_54
40. Vrigkas M, Nikou C, Kakadiaris I, A review of human activity recognition methods. Frontiers in Robotics and AI. 2015, 2,1-28. https://doi.org/10.3389/frobt.2015.00028.
41. Wenjuan G, Xuena Z, Jordi G, Andrews S, Thierry B, Changhe T, El-hadi Z. Human Pose Estimation from Monocular Images: A Comprehensive Survey, Sensors (Basel, Switzerland). 2016, 16, 1-39. DOI: 10.3390/s16121966.
42. Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding. 2006, 104(2-3),249-257. https://doi.org/10.1016/j.cviu.2006.07.013.
43. Xu K, Qin Z, Wang G. Recognize human activities from multi-part missing videos. IEEE International Conference on Multimedia and Expo. 2016, 976–990. DOI: 10.1109/ICME.2016.7552941.

44. Yadav S.K, Singh A, Gupta A, Real-time Yoga recognition using deep learning. *Neuralomputing and Applications. 2019, 9349– 9361.DOI: 10.1007/s00521-019-04232-7*