International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206623

Yoga Pose Detection and Classification Using Deep Learning

Deepak Kumar, Anurag Sinha
Department of Information Technology, Research Scholar, Amity University, Jharkhand, India

Article Info
Volume 6, Issue 6
Page Number: 160-184
Publication Issue :
November-December-2020

Article History
Accepted : 15 Nov 2020
Published : 28 Nov 2020

ABSTRACT

Yoga is an ancient science and discipline originated in India 5000 years ago. It is used to bring harmony to both body and mind with the help of asana, meditation and various other breathing techniques It bring peace to the mind. Due to increase of stress in the modern lifestyle, yoga has become popular throughout the world. There are various ways through which one can learn yoga. Yoga can be learnt by attending classes at a yoga centre or through home tutoring. It can also be self-learnt with the help of books and videos. Most people prefer self-learning but it is hard for them to find incorrect parts of their yoga poses by themselves. Using the system, the user can select the pose that he/she wishes to practice. He/she can then upload a photo of themselves doing the pose. The pose of the user is compared with the pose of the expert and difference in angles of various body joints is calculated. Based on thisdifference of angles feedback is provided to the user so that he/she can improve the pose.

Keywords : Pose, Self-Learning, Posenet, Deep Learning, Pose Classification.

I. INTRODUCTION

Human posture assessment is a difficult issue in the control of PC vision. It manages confinement of human joints in a picture or video to shape a skeletal portrayal. To consequently recognize an individual's posture in a picture is a troublesome errand as it relies upon various perspectives, for example, scale and goal of the picture, enlightenment variety, foundation mess, dress varieties, environmental factors, and connection of people with the environmental factors [1]. An utilization of posture assessment which has pulled in numerous analysts in this field is practice and wellness. One type of activity with multifaceted stances is yoga which is a deep rooted practice that begun in India however is presently celebrated overall due to its numerous profound, physical and mental benefits [2].

The issue with yoga anyway is that, much the same as some other exercise, it is of most extreme significance to rehearse it accurately as any erroneous stance during a yoga meeting can be ineffective and conceivably inconvenient. This prompts the need of having a teacher to manage the meeting and right the person's stance. Since not all clients approach or assets to a teacher, a computerized reasoning based application may be utilized to recognize yoga presents and give customized input to assist people with improving their structure [2].

Lately, human posture assessment has profited extraordinarily from profound learning and gigantic gains in execution have been accomplished [3]. Profound learning approaches give a more clear method of planning the structure as opposed to managing the conditions between structures physically. [4] utilized profound figuring out how to distinguish 5 exercise presents: pull up, swiss ball hamstring twist, push up, cycling and strolling. Nonetheless, utilizing this technique for yoga presents is a moderately more up to date application [2].

This undertaking centers around investigating the various methodologies for yoga present order and looks to accomplish knowledge into the accompanying: What is present assessment? What is profound realizing? How can profound learning be applied to yoga present order continuously? This task utilizes references from meeting procedures, distributed papers, specialized reports and diaries. Fig. 1 gives a graphical review of points this paper covers. The main part of the venture discusses the history and significance of yoga. The subsequent segment discusses present assessment and clarifies various kinds of posture assessment strategies in detail and goes one level further to clarify discriminative strategies – learning based (profound learning) and model. Diverse posture extraction strategies are then talked about alongside profound learning based models - Convolutional Neural Organizations (CNNs) and Recurrent Neural Networks (RNNs).

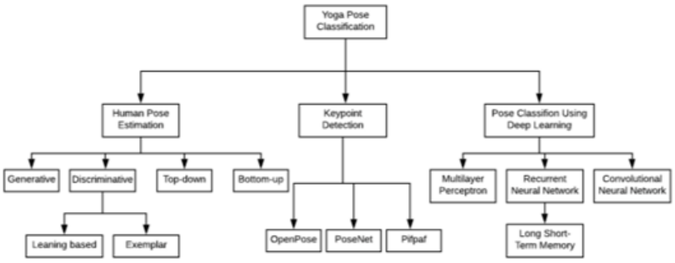


Fig. 1. Conceptual outline of topics

II. HISTORY

People are inclined to musculoskeletal problems with maturing and mishaps [5]. So as to forestall this a few, type of actual exercise is required. Yoga, which is a physical and otherworldly work out, has increased colossal centrality in the network of clinical scientists. Yoga has the capacity to totally fix illnesses with no prescriptions and improve physical and mental wellbeing [6]. A tremendous assemblage of writing on the clinical uses of yoga has been created which incorporates positive self-perception mediation, heart restoration, psychological sickness and so forth [6]. Yoga contains different asanas which speak to actual static stances. The utilization of posture assessment for yoga is trying as it includes complex setup of stances. Moreover, some best in class strategies neglect to perform well when the asana includes even body act or then again when both the legs cover one another. Subsequently, the need to build up a strong model which can help advocate self-taught yoga frameworks emerges

III. HUMAN POSE ESTIMATION

Human stance acknowledgment has made colossal headways in the previous years. It has advanced from 2D to 3D present assessment and from single individual to multi individual posture assessment. [16] employments present assessment to fabricate an AI application that identifies shoplifters while [17] utilizes a solitary RGB camera to catch 3D stances of numerous individuals continuously. Human posture assessment calculations can be broadly coordinated in two different ways. Calculations prototyping assessment of human stances as a mathematical estimation are named generative strategies while calculations demonstrating human posture assessment as a picture preparing issue are named discriminative strategies [7]. Another method of grouping these calculations depends on their strategy for or king. Calculations beginning from a more significant level speculation and descending are called top-down strategies, while calculations that start with pixels and move upwards are rung base strategies [8].

A. GENERATIVE

Generative methods give a procedure to foresee the highlights from a given posture speculation. They start with instating the stance of the human body and venture it to the picture plane. Changes are made to make the extended picture and current picture perceptions consistent. Generative based methodologies offer simple speculation because of less requirement of a preparing present dataset [9]. Nonetheless, because of the high dimensional projection space search, this technique isn't considered computationally plausible, and is in this manner more slow when contrasted with discriminative techniques. [10] portrays a generative Bayesian technique to follow 3D fragmented human body figures in recordings. This is a probabilistic technique which comprises of a generative model for picture appearance, an underlying likelihood dissemination over joint points and represent that speaks to development of people and a strong probability work. Despite the fact that the strategy can follow people in obscure convoluted foundations, it faces the danger of in the long run forgetting about the object.

B. DISCRIMINATIVE

In opposition to generative techniques, discriminative strategies start with the proof of the picture and get familiar with a strategy to demonstrate the connection between the human postures and proof on the premise of preparing information. Model testing in discriminative strategies is much quicker instead of generative strategies because of the hunt in an obliged space rather than a high dimensional include space [7]. [11] investigates a discriminative based learning technique to get 3D human posture from outlines. This methodology doesn't need a body model unequivocally nor any earlier marked portions of the body in the picture. It reestablishes the posture utilizing non-straight relapse dependent on the shape descriptor vectors brought naturally from outlines of pictures. It utilizes Relevance Vector Machine (RVM) regressors and damped least squares for relapse [11]. The strategy, however expanding the precision by multiple times, isn't sufficiently exact, as there are a few examples of erroneous postures and results indicating critical fleeting jitter. Discriminative strategies are further ordered into learning techniques and model strategies [7].

I. LEARNING BASED – DEEP LEARNING:

One significant learning-based technique is profound realizing which is based upon Artificial Neural Organizations (ANNs). ANN is similar to the human cerebrum where the units in an ANN speak to the neurons in the human mind, and loads speak to the quality of association between neurons.Profound learning gives a start to finish design that permits programmed learning of key data from pictures. One famous profound learning model which has been generally utilized for present assessment is Convolutional Neural Network (CNN) which will be talked about later. [20] have added to the exploration by utilizing CNNs and stacked auto-encoder calculations (SAE) for distinguishing yoga stances and Indian old style move structures. In any case, their presentation assessment is done distinctly on pictures and not on recordings.

II. EXEMPLAR METHODS:

In model techniques, present assessment depends on a special arrangement of postures with their equal portrayals [7]. Characterization calculations, for example, arbitrary timberlands and randomized trees are strong and quick enough to deal with this. Irregular woods is comprised of different randomized choice trees and is henceforth called a group classifier. It comprises of non-terminal hubs which have a choice capacity to foresee the similitudes in pictures.

Fig. 8 gives a model. [19] utilized an improved variant of irregular woodlands which contained two levels of arbitrary backwoods. The first layer of the tree went about as a discriminative classifier to arrange body parts and passed the arrangement results to the second layer which anticipated joint areas in the body. Another approach like irregular timberlands is Hough backwoods [7] which comprises of choice woods blends, where the terminal hub in each tree is either a relapse or arrangement hub. Upgraded Hough trees have a sections objective ("PARTS") which is an advanced target based on discrete data gain.

C. TOP- DOWN

Most examinations allude to generative strategies as top-down techniques [7]. The top-down methodology acquires keypoints by utilizing a module to identify human subjects to which a posture assessor can be applied. The essential bit of leeway of top-down techniques is their capacity to separate the undertaking into various more modest assignments. These more modest errands incorporate distinguishing the item followed by present assessment. For this, the identifier should be ground-breaking enough to recognize hard or relatively more modest items with the goal that the exhibition of the posture assessor is improved. [8] utilizes a top down way to deal with articulate human posture assessment and following. Their top down methodology includes three segments – a human applicant indicator, a solitary individual posture assessor and a human posture tracker [8]. An overall article identifier is picked to recognize human applicants after which a fell pyramid tracker is utilized to perceive the identical human posture and finally, a stream based posture tracker is utilized to relegate a particular and transiently reliable id to every human possibility for multi present following. Despite the fact that this examination builds up a secluded framework to human posture assessment and following, they accomplish a normal accuracy of just 69.4 for present assessment and 68.9 for present following, leaving a great deal of extension for development.

D. BOTTOM – UP:

This methodology includes discovery of human keypoints from likely subjects and putting together them into human appendages utilizing a few information affiliation instruments. The expense of calculation in this technique doesn't rely upon the quantity of human subjects in the pictures. In this manner, it gives an awesome remuneration among cost and exactness. A few studies allude to base up techniques as discriminative strategies [7]. [12] proposes a novel methodology that consolidates customary base up and top-down strategies for multi-individual posture assessment. The feed sending to the organization is done in a base up style while parsing of postures alongside bouncing box limitations is acted in a top-down design. The highlights from the picture are connections of network between the joints. This emaining organization is only ResNet50, which is a profound neural organization design. In spite of the fact that this examination doesn't zero in on arranging the stances, it shows how profound learning designs like ResNet50 can be utilized so as to make human posture assessment exact.

IV. KEYPOINT DETECTION METHODS

A. OPENPOSE:

OpenPose is a multi-individual continuous keypoint location which acquired a transformation the field of posture assessment. It was developed in Carnegie Mellon University (CMU) by the Perceptual Registering Lab[13]. It utilizes CNN based engineering to recognize facial, hand and foot keypoints of a human body from single pictures. OpenPose recognizes human body joints utilizing a RGB camera. OpenPose keypoints incorporate eyes, ears, neck, nose, elbows, shoulders, knees, wrists, lower legs and hips. It presents the outcomes acquired by handling contributions from a camera continuously or pre-recorded recordings or static pictures as 18 basic keypoints. Along these lines, it discovers its utilization in a assortment of utilizations going from sports, reconnaissance, action location to yoga present acknowledgment. The work proposed in [18] utilizes OpenPose for introductory keypoint recognizable proof followed by CNN for characterization of yoga presents. Notwithstanding, they accomplish a precision of just 78% which could be because of the restricted dataset they utilized or design and hyperparameter tuning of their CNN model.

The primary stage in OpenPose is identifying keypoints of each individual in the picture which is followed by relegating parts to each particular person. Fig. 2 portrays the engineering of the OpenPose model [13]. OpenPose network begins with extraction of highlights from the picture utilizing the underlying layers (VGG - 19 as appeared in Fig. 2). These highlights are then passed to two convolutional layer branches which run in equal. A forecast of 18 certainty maps, which speaks to explicit portions of the human body, is made by the principal branch. Then again, 38 Part Affinity Fields (PAF) which indicate the affiliation degree between parts is anticipated continuously branch. More stages are utilized to make refinement to the expectations produced using the past branch. Bipartite charts are framed between various parts utilizing part certainty maps. The connections which are more fragile in these diagrams are taken out utilizing the PAF esteems. With these means, human skeletons are assessed for each individual in the edge or picture.

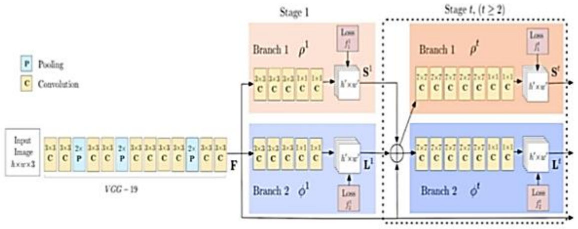


Fig. 2. OpenPose architecture

B. POSENET:

PoseNet is another profound learning structure like OpenPose which is utilized for ID of human postures in pictures or video successions by distinguishing joint areas in a human body. These joint areas or keypoints are listed by "Part ID" which is a certainty score whose worth lies in the scope of 0.0 and 1.0 with 1.0 being the best. The PoseNet model's execution shifts relying upon the gadget and yield step [14]. The PoseNet model is invariant to the size of the picture, consequently it can anticipate present situations in the size of the genuine picture regardless of whether the picture has been downscaled.

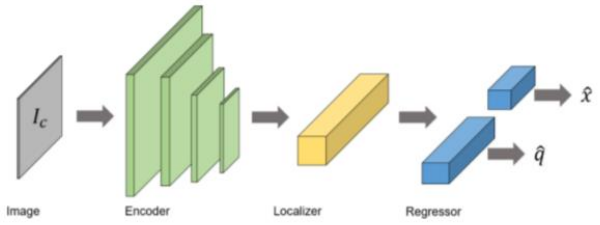


Fig. 3. System architecture of PoseNet

C. PIFPAF:

PifPaf is another technique dependent on the base up approach for 2D multi-individual human posture assessment. It utilizes a Part Intensity Field (PIF) for body part limitation and a Part Association Field (PAF) for relationship of body parts to shape full human stances [15]. The model beats other techniques regarding a lower goal and better execution in stuffed places principally due to the accompanying: (a) fine data encoded in a more current composite field PAF, (b) the determination of Laplace misfortune that incorporates an assessment of vulnerability. The model engineering settles upon a totally convolutional sans box plan [15].

Fig. 4 speaks to the engineering of PifPaf [15]. The information picture is of size (H, W). It has the RGB channels which is appeared by 'x3'. The encoder depends on neural organizations, and it produces the PIF field with 17 x 5 channels and PAF field with 19×7 channels. '//2' speaks to an activity with steps of 2. The PIF and PAF fields are changed over by the decoder into present directions which have 17 joints each. Each joint is a 2D portrayal and has X and Y organizes alongside a certainty score.

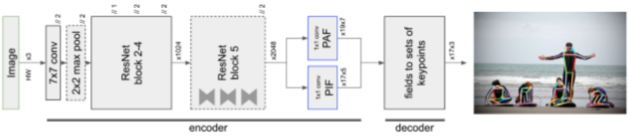


Fig. 4. Pifpaf architecture

This model engineering is a ResNet based organization. The certainty, careful area and joint size is anticipated by one of the head networks which is called PIF, and the relationship between various parts are anticipated by the other head network which is PAF. The strategy is thus called PifPaf.

V. POSE CLASSIFICATION USING DEEP LEARNING

Profound learning is generally utilized for picture order undertakings wherein the model takes input as pictures and yields a forecast. Profound learning calculations utilize neural organizations to decide the association between the information and yield. For present assessment issues, the picture with posture of people is taken as information and the profound learning model attempts to adapt effectively the various postures in order to precisely group them. As one can figure, this could be a computationally costly assignment if the quantity of pictures is enormous. Additionally, as we need precise outcomes we would not need to settle on the nature of the pictures as that could influence the highlights removed by the model. Consequently, in this venture we propose utilizing OpenPose (a pretrained model) to separate keypoints of the human joint areas from the pictures and afterward preparing the profound learning model on these keypoints. The following are some essential profound learning models utilized for characterization issues.

A. MULTILAYER PERCEPTRON (MLP)

MLP is an old style neural organization that has one info and one yield layer. The transitional layers between the info and yield layer are known as concealed layers. There can be at least one shrouded layers. MLPs structure a completely associated network as each hub in one layer has an association to each hub in another layer. A completely associated network is an establishment for profound learning. MLP is well known for directed characterization where the information is relegated a mark or class. [21] employments MLP for human posture order by separating keypoints from low goal pictures utilizing Kinect sensor.

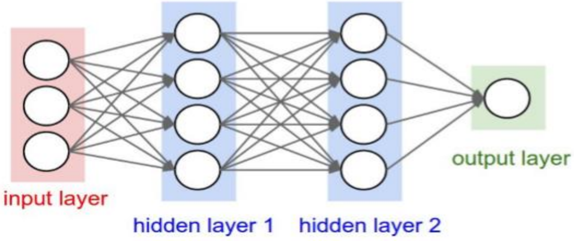


Fig. 5. Schematic diagram of multilayer perceptron

B. Repetitive NEURAL NETWORK (RNN)

RNNs are neural organization designs that are utilized for grouping expectation issues. Grouping expectation issues can be one to many, numerous to one, or numerous to many. In RNNs, the past information of a neuron is saved which helps in dealing with the successive information. Thus, the setting is protected, and yield is created considering the recently learned information. RNNs are most regularly utilized for normal language handling (NLP) issues where the information is normally demonstrated in groupings.

Be that as it may, in action acknowledgment or posture arrangement undertakings as well, there is a reliance between the recently performed activity and the following activity. In the event of yoga also, the unique situation or on the other hand data of introductory or go-between presents is significant in anticipating the last posture. Yoga can in this way be considered as a succession of postures. This settles on RNNs an appropriate decision for yoga present arrangement as consecutive assessment of joint areas can have the option to more readily catch the reliance between joint areas. For a similar explanation, [22] use RNN for human posture assessment.

The issue with RNNs anyway is that they can't protect long haul conditions. Now and then, late data is adequate to direct the current undertaking, yet there are situations when the hole between the applicable data and the current assignment turns out to be excessively huge.

In such cases RNNs fizzle as they can't interface this data. With regards to yoga, if the delegate steps in a yoga asana are too much, RNNs think that its hard to monitor the starting advances which are required so as to anticipate the current errand. This issue is called as the long haul reliance issue.

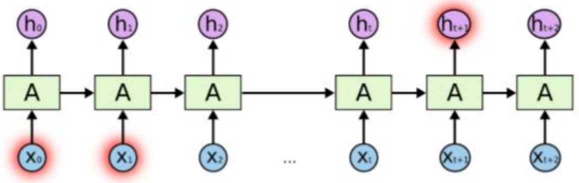


Fig. 6. Long-term dependency in RNN

I. LONG SHORT-TERM MEMORY (LSTM):

So as to manage the above long haul reliance issue, an extraordinary kind of RNN exists which is called LSTM. A LSTM is a renowned RNN that can undoubtedly recall data or then again information for a considerable length of time timeframes which is its default conduct. The key thought which makes this conceivable is cell state. A cell state permits unaltered data stream. It tends to be considered as a transport line. LSTMs can add and kill information from the cell state utilizing administrative structures known as entryways. These exceptional entryways consider alternatively letting the data through. LSTM utilizes three entryways, specifically information, refresh and overlook. A LSTM can accordingly specifically overlook or recollect the learnings. As LSTMs take into account longer maintenance of the info state in the organization, they can proficiently deal with long successions and give great outcomes. [23] discusses how LSTM can be utilized with CNN for human action acknowledgment to accomplish high exactness.

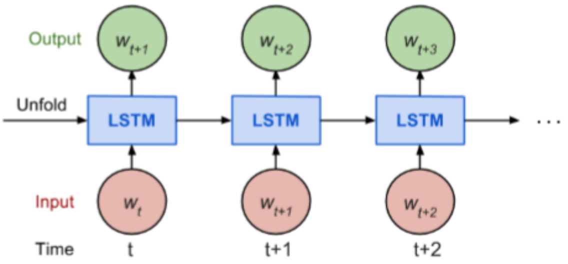


Fig. 7. LSTM architecture

D. CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is a kind of neural organization which is broadly utilized in the PC vision space. It has end up being exceptionally successful with the end goal that it has become the go-to strategy for most picture information. CNNs comprise of at least one convolutional layer which is the main layer and is mindful for include extraction from the picture. CNNs perform include extraction utilizing convolutional channels on the information and examining a few pieces of the contribution at a given time prior to sending the yield to the ensuing layer. The convolutional layer, using convolutional channels, creates what is known as an element map. With the assistance of a pooling layer, the dimensionality is decreased, which decreases the preparation time and forestalls overfitting. The most well-known pooling layer utilized is max pooling, which takes the most extreme incentive in the pooling window.

CNNs show an extraordinary guarantee in present characterization assignments, in this way making it an exceptionally attractive decision. They can be prepared on keypoints of joint areas of the human skeleton or can be prepared legitimately on the pictures. [4] utilized CNN to distinguish human postures from 2D human exercise pictures and accomplished a precision of 83%. Then again, [18] utilized CNN on OpenPose keypoints to arrange yoga presents and accomplished a precision of 78%. Despite the fact that, the exactness isn't actually similar as the dataset alongside the CNN engineering and activities being ordered are extraordinary, [18] shows how utilizing CNNs on OpenPose keypoints merits investigating.

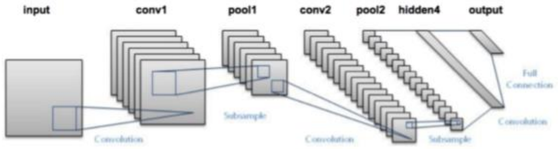


Fig. 8. CNN architecture layers

On account of keypoints, CNN removes highlights from 2D directions of the OpenPose keypoints utilizing the equivalent convolutional channel strategy clarified previously. In light of the channel size, the convolutional channel slides to the following arrangement of information. After the convolution, an initiation work Redressed Linear Unit (ReLU) is commonly applied to add nonlinearity in the CNN, as the genuine world information is generally nonlinear and the convolution activity without anyone else is straight. Tanh and sigmoid are other enactment capacities, yet ReLU is generally utilized due to its better exhibition.

VI. SUMMARY OF CURRENT STATE OF THE START

A great deal of work has been done in the past in building frameworks that are mechanized or semiautomated which help to investigate exercise and sports exercises, for example, swimming [24], ball [25] and so on Patil et al. [26], proposed a framework for distinguishing yoga act contrasts between a specialist and a professional utilizing speeded up vigorous highlights (SURF) which employments data of picture shapes. Notwithstanding, depicting and contrasting the stances nearly by utilizing just the shape data isn't adequate.

A framework for yoga preparing has been proposed by Luo et al. [27] which comprises of inertial estimation units (IMUs) and tactors. Be that as it may, this can be awkward to the client and at the equivalent time influence the common yoga present. [28] introduced a framework for yoga present location for six postures utilizing Adaboost classifier and Kinect sensors and accomplished a precision of 94.8%. Notwithstanding, they have utilized a profundity sensor based camera that may not be consistently available to clients. Another framework for yoga present adjustment utilizing Kinect has been introduced by [29] which considers three yoga presents, hero III, descending canine and tree present. Nonetheless, their outcomes are not very great, and their exactness score is just 82.84%. The customary strategy for skeletonization has now been supplanted by profound learning-based techniques.

Deep learning is a promising space where a ton of exploration is being done, empowering us to dissect gigantic information in an adaptable way. When contrasted with conventional AI models where highlight extraction and designing is an unquestionable requirement, profound learning kills the need to do as such by understanding complex examples in the information and removing highlights all alone.

VII. 7. HYPOTHESIS

A profound learning model for arranging yoga stances can be fabricated where the underlying keypoint extraction of the human joint areas is finished utilizing OpenPose. The model can fuse highlight extraction capacities of CNN alongside setting maintenance capacities of LSTM to adequately arrange yoga presents in prerecorded recordings and furthermore continuously [2]. This model can be considered as a half and half model.

We additionally plan to explore different avenues regarding essential CNN organizations and contrast the exhibition and the half breed model. Kinect sensors could be an approach to perform human posture assessment, however it accounts for extra gear and concentrated equipment and the exhibition isn't in every case great in diverse environmental factors. AI models, despite the fact that not broadly utilized for human posture assessment, will be investigated for correlation with the profound learning models. The assessment of the yoga present arrangement framework will be finished by utilizing grouping scores, disarray network and assessments by individuals. The framework will anticipate the yoga present arrangement being performed by the client continuously and we can inspect if the expectation made by the framework is right. The outcomes will likewise be contrasted with existing strategies.

VIII. 8. EVALUATION METRICS

A. Classification Score:

Arrangement score alludes to what we typically mean by exactness of the model. It very well may be portrayed as the extent of number of forecasts that were right to the complete information tests.

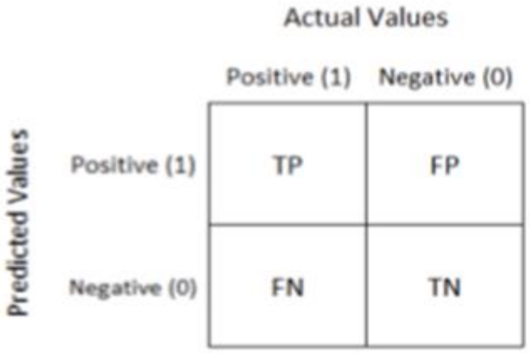
In the event of multiclass order, this measurement gives great outcomes when the quantity of tests in each class are nearly the equivalent.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

B. Disarray framework

Disarray framework speaks to a lattice which clarifies the precision of the model totally. There are four significant terms with regards to estimating the presentation of a model.

- ✓ True Positive: Predicted esteem and the genuine yield are both 1.
- ✓ True Negative: Predicted esteem and the genuine yield are both 0.
- ✓ False Positive: Predicted esteem is 1 yet the genuine yield is 0.
- ✓ False Negative: Predicted esteem is 0 yet the genuine yield is 1.



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

Fig. 9. Sample confusion matrix

Fig. 9. shows an essential disarray lattice for twofold grouping. The askew qualities speak to the examples that are accurately characterized and along these lines, we generally need the askew of the framework to contain the most extreme worth. If there should arise an occurrence of a multiclass grouping, each class speaks to one column what's more, segment of the framework.

C. Model accuracy and model loss curve:

These bends are additionally alluded to as expectations to absorb information and are generally utilized for models that learn gradually after some time, for instance, neural organizations. They speak to the assessment on the preparing and approval information which gives us a thought of how well the model is learning and how well is it summing up. The model misfortune bend speaks to a limiting score (misfortune), which implies that a lower score brings about better model execution. The model precision bend speaks to a expanding score (exactness), which implies that a higher score signifies better execution of the model. A decent fitting model misfortune bend is one in which the preparation and approval misfortune decline also, arrive at a state of soundness and have negligible hole between the last misfortune esteems. On the other hand, a decent fitting model exactness bend is one in which the preparation and approval precision increment and become stable and there is a base hole between the last exactness esteems.

IX. LITERATURE REVIEW

AI strategies may maybe rely significantly upon heuristic human element extraction in everyday errands of location of social exercises. It is limited by human zone mindfulness. To talk this danger, creators have decided on a couple of techniques like profound learning methods. These strategies could consequently extricate explicit highlights during the reparing stage from crude sensor information, and afterward low-level fleeting qualities with ignificant level unique requests would be introduced. With regards to the freed application from profound learning approaches in fields like picture grouping, voice acknowledgment, preparing of regular language, and some others, it has been developing into a novel examination way in design location and to move it to an area of human action identification. Table 2 shows a couple of the current AI alongside the exactness of HAR. Here, we have referenced just the techniques that give the best exactness for the most extreme number of subjects. For instance, on the off chance that the subject tally is less, at that point the precision might be better.

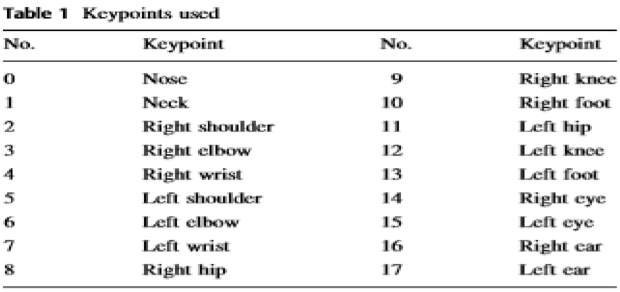


Table 1 Keypoints used

No.	Keypoint	No.	Keypoint
0	Nose	9	Right knee
1	Neck	10	Right foot
2	Right shoulder	11	Left hip
3	Right elbow	12	Left knee
4	Right wrist	13	Left foot
5	Left shoulder	14	Right eye
6	Left elbow	15	Left eye
7	Left wrist	16	Right ear
8	Right hip	17	Left ear

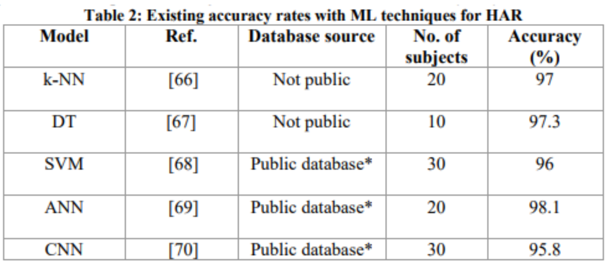


Table 2: Existing accuracy rates with ML techniques for HAR

Model	Ref.	Database source	No. of subjects	Accuracy (%)
k-NN	[66]	Not public	20	97
DT	[67]	Not public	10	97.3
SVM	[68]	Public database*	30	96
ANN	[69]	Public database*	20	98.1
CNN	[70]	Public database*	30	95.8

Chen et al. [30] proposed a system for distinguishing a yoga act utilizing a Kinect camera. Those assembled an amount of 300 recordings of 12 yoga stances from 5 yoga pros with each present performed on five events. At first, the closer view part is fragmented from the cut, and the star skeleton is used and acquired an exactness of 99.33%. The creators in [31] proposed a stance location system utilizing a quality Kinect camera with a goal of 640 X 480. They saw six postures performed by five volunteers. They separated 21000 casings from the Kinect camera by utilizing a foundation deduction technique with a 74% precision. Later the creators in [32] proposed another posture discovery methodology utilizing Kinect camera. Also, Wang et al. [33] proposed a position acknowledgment methodology using the Kinect camera. They isolated the human blueprint and utilized a learning vector quantization neural framework for five fundamental stances. The system accuracy was roughly 98 %. onetheless, these outcomes have high precision rates, and they are seen as security prominent.

Yao et al. [34] have proposed a human stance recognizable proof system using a disconnected RFID signal. This methodology sees the human postures subordinate over the assessment of RSSI signal models, which made while singular plays out the posture in a RFID name group and a RFID gathering mechanical assembly. An accuracy of 99 percent for 12 stances was achieved through the structure. In spite of its higher exactness rate, the RSSI signal is climate subordinate; likewise, the foundation and upkeep of this technique are modern; an immediate aftereffect of its various parts. In [35], the creators have developed a savvy petition tangle that sees four represents that are seen during supplication. The tangle has a couple power identifying resistive strips inside and perceives the region where the individual's body is crushing. The tangle apparent the stances assembled from 30 individuals with 100% exactness. Regardless, such a procedure can't recognize the body parts if they are not pushed on the tangle. An earlier variation [36] of this paper grasps comparative hardware, including three tomahawks (x, y, and z) hubs, for the security protected posture acknowledgment framework. Exploratory results achieved a high ordinary F1-score of roughly 98% in various blends of measurements. In any case, the outcomes rely upon only six postures assembled from four volunteers.

In [37], the creators utilized an inclination histogram, and Fourier descriptors subject to centroid highlights are used. By then, Jain et al. [37] used two classifiers, SVM also, K-NN, to see the activities of two open datasets. They utilized six inertial assessment units to construct the system. The creators used the Random Forest classifier to portray the exercises. Finally, an overall accuracy of 84.6% was refined. The creators in [38] proposed a classifier by coordinating the profound CNN and LSTM for grouping hand movements of 5 developments with a F1 score of 0.93 and 0.95 for the two classifiers independently. Lin et al. [39] presented another iterative CNN approach with autocorrelation pre-preparing, as opposed to ordinary little scope Doppler preprocessing, which can absolutely arrange seven activities or five subjects. Furthermore, this framework used an iterative profound learning structure to normally characterize and separate highlights. Finally, traditional classifiers were utilized to check different activities subordinate upon input radar signals. In spite of the fact that the above models could, overall, see human exercises, the overall framework structure is tolerably perplexing. A lightweight profound learning model is proposed by [40] for HAR and sent it on Raspberry Pi3. This example was made using a shallow RNN in mix with LSTM, and its overall precision upon the WISDM dataset got around 96%. Despite the fact that this model has high exactness with great plan, it was simply evaluated on a solitary dataset, which has just six activities, that doesn't exhibit that the suggested plan has extraordinary speculation capacity. In [41] presented a profound learning configuration named Inno-HAR by incorporating commencement NN with RNN for movement grouping. The creators used separate convolution to dislodge the ordinary convolution, which achieved the target of reducing model boundaries. The results showed an awesome effect. Notwithstanding, the model met scarcely, causing a huge load of time to be squandered at preparing [42].

Countless these datasets are assembled from the sources, for instance, on the web accounts, films, pictures, sports chronicles, etc. Some of them give rich imprint information; in any case, they need human stance variety. Chen et al. [43], starting late observed that the pictures considered for remarks are of high type with enormous target things. For occurrence, in the MPII dataset, around 70% of the photos includes person things with height in excess of 250 pixels. Thusly, missing a ton of grouped assortment in individual positions and target object size in these datasets, they can't meet the first rate necessities of uses, for instance, direct assessment. Barely any works have been done on yoga act arrangement for applications, for example, self-getting ready [44]. Anyway, these works incorporate a yoga dataset with a less number of pictures or accounts and try not to consider the gigantic collection of positions. In this way, they need theory and are far from complex yoga present arrangement. The creators in , applied posture arrangement for traditional move and Yoga presents utilizing CNN and SAE calculation. Nevertheless, they have evaluated their outcomes on static pictures and not on accounts. OpenPose is a continuous framework introduced by the Perceptual Figuring Lab of Carnegie Mellon University (CMU) to commonly recognize a human body, hand, facial, and foot key focuses on single pictures. It is critical with regards to act acknowledgment and gives the human body joint regions using convolutional neural frameworks (CNNs).

The creators in propose a clever structure prepared for seeing six postures for learning Yoga that can catch up to 6 people all the while with intelligent Kinect gadget. It is moreover consolidated with request voices to envision the rules and pictures about the postures. For acknowledgment, the structure used the AdaBoost calculation. The info information was set up by an expert yoga guide and acquired an exactness of 94.58%. The creators in proposed an IoT-based system named as protection saving yoga present discovery structure, using the Deep CNN (DCNN) for inferior quality sensor pictures. They utilized WSN, which comprises of the three hubs, sensor module, Wi-Fi module to associate the worker. They gathered the information from 18 volunteers on 26 distinctive yoga stances with the video length of 20 seconds. From the video document, they produced the pictures and applied DCNN models on it. The F1 score of this technique in the wake of utilizing ten times cross-approval methods is about 0.99 and 0.98 with 3-tomahawks and just a single hub (y-pivot), individually. The worker side strategy of their work is appeared in Figure 4. The three various casings in three bearings are consolidated, and the profound CNN is applied to the picture for characterization to perceive the stance in the given picture .

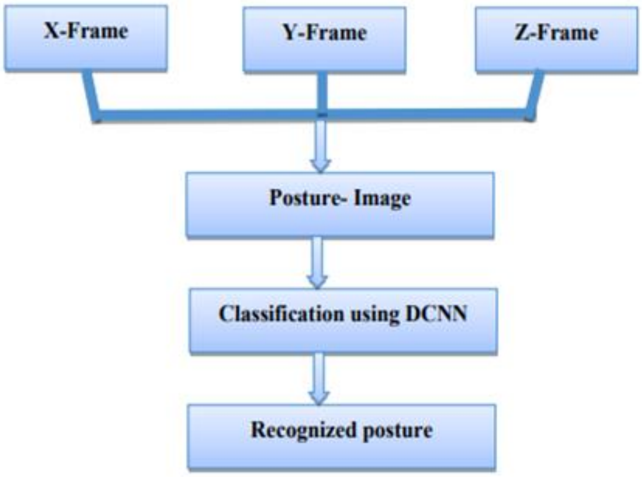
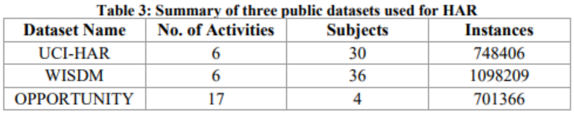


Figure 10: server-side classification strategy

In [45], the creators proposed a yoga act acknowledgment framework utilizing a 2-stage classifier (i.e., BackPropagation Artificial Neural Network (BP-ANN) and Fuzzy C-implies) along with assessment measurements to direct the students. As a first classifier, they utilized BP-ANN to partition the yoga stances into various classifications, and as a second classifier fluffy Cmeans classifier to order the places of various kinds. To assess the technique, they utilized a wearable gadget with 11 IMUs and a stance information base with 11 subjects, roughly 211 a great many information outlines with 1800 stance cases. 30% of data used to prepare, and the remainder of the information utilized as test information, and the outcome was 95.39% precision. There exist numerous works prior to utilizing this 2-stage arrangement by utilizing procedures like HMM, SVM, and Decision tree, and so on For example, Kang et al. [46] applied DT for pose acknowledgment utilizing portrayal highlights utilizing the body skeleton model. Despite the fact that the registering cost isn't more, the accuracy was tolerably low.Wu et al. [47] proposed three measures, specifically joint point, arm heading, and regular advancement type, which could be used generally speaking to evaluate the lower arm and upper arm. The creators in proposed a methodology for yoga pose examination using present acknowledgment. As per this method at first perceives a stance utilizing OpenPose also, a camera. By then, it registers separation between the body points between a teacher and the client. In case it is greater than the given limit, the technique proposes the amendment of the part. With this suggestion, it is typical that people can practice Yoga wherever, including home. Thusly, everyone can rehearse Yoga, paying little heed to mature enough or prosperity. For assessments, the creators applied the suggestion to 4 particular circumstances, for instance, unique body sizes, different stature, different ages, and distinctive camera division, with three Yoga presents. Prior to this work, the creators in, presented video demonstrating and input. They passed on bearings or self-checked contributions to understudies to improve conduct changes. The inspiration driving this assessment was to review video self-evaluations. The exactness was dictated by apportioning the quantity of right walks through the general number of ventures by investigating the errand. The creators in , proposed a profound neural organization, incorporates CNN with LSTM approach. The highlights are extricated naturally, and order was performed with two or three model boundaries. LSTM is a variety of the RNN, that is progressively proper for handling fleeting groupings. Concurring to this strategy, the information is gathered from sensors and will be given to the 2-layer LSTM, alongside convolutional layers. Worldwide Average Pooling layer and Batch Normalization likewise applied to accelerate the cycle and to accomplish better results. To exhibit the generalizability and ampleness of their work, they used three public datasets (UC-HAR, WISDM, and OPPORTUNITY), and a rundown of these datasets is appeared in Table 3. In expansion to the exactness, they additionally considered the F1 score for assessment of the model what's more, accomplished roughly 96%, 96%, and 93% for these three datasets, separately.

Table 3: Summary of three public datasets used for HAR



Dataset Name	No. of Activities	Subjects	Instances
UCI-HAR	6	30	748406
WISDM	6	36	1098209
OPPORTUNITY	17	4	701366

The creators in proposed a technique to perceive the Yoga presents precisely utilizing profound learning technique. For the investigation, they considered the open dataset, which comprises of a bunch of six yoga asanas from 15 volunteers (10 male, five female) with the norm camera. The creators utilized a mixture CNN (utilized for include extraction) alongside the LSTM approach (utilized for fleeting forecasts) to perceive the stances of Yoga on recordings and accomplished a precision of 99% around. They additionally thought about the work in ongoing with 12 various people (five guys, seven females) and acquired almost 98% precision. The central issues are distinguished by the OpenPose are indicated graphically in Figure 5, and similar basic focuses used in the human body are spoken to in Table 4.

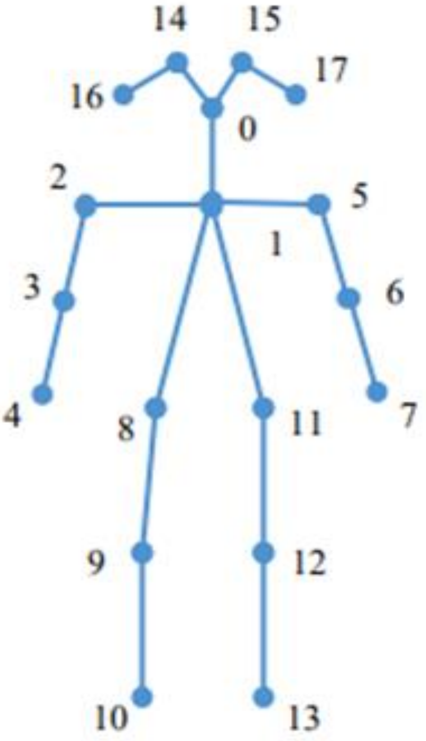
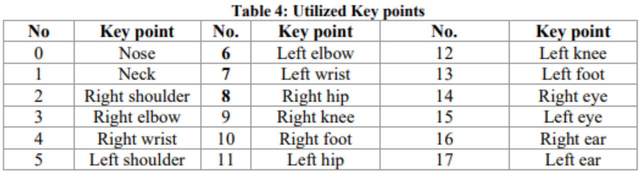


Figure 11. Key point in the human body detection by openPose

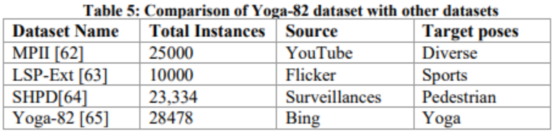
The current datasets for present acknowledgment worse as far as variety, impediment, and perspectives. As such, present acknowledgment restricted on less number of postures. The creators in [48], proposed a thought of fine-grained present characterization, in which they figured the act acknowledgment like not as a basic grouping, but rather it is fine-grained. The creators moreover proposed a dataset named Yoga-82, for enormous scope grouping with 82 classes. Not at all like with different datasets, Yoga-82 planned expressly for Yoga presents with various varieties of the stances.

Table 4: Utilized Key points



No	Key point	No.	Key point	No.	Key point
0	Nose	**6**	Left elbow	12	Left knee
1	Neck	**7**	Left wrist	13	Left foot
2	Right shoulder	**8**	Right hip	14	Right eye
3	Right elbow	9	Right knee	15	Left eye
4	Right wrist	10	Right foot	16	Right ear
5	Left shoulder	11	Left hip	17	Left ear

Table 5: Comparison of Yoga-82 dataset with other datasets



Dataset Name	Total Instances	Source	Target poses
MPII [62]	25000	YouTube	Diverse
LSP-Ext [63]	10000	Flicker	Sports
SHPD[64]	23,334	Surveillances	Pedestrian
Yoga-82 [65]	28478	Bing	Yoga

The examination of this dataset with different datasets is appeared in Table 5. The yoga-82 dataset contains three levels, including body positions, varieties, and names of the stances. The commitments of their work incorporate a fine-grained characterization with a huge scale dataset and assessed the exhibition with CNN. They additionally adjusted the engineering of Dense Net to utilize the chain of importance of the dataset and to acquire the exact present acknowledgment results.

X. Research Methodology

Our methodology expects to naturally perceive the client's Yoga asanas from ongoing and recorded recordings. The technique can be decayed into four fundamental advances. To begin with, information assortment is performed which can either be an ongoing cycle running in corresponding with identification or can be recently recorded recordings. Second, OpenPose is utilized to distinguish the joint areas utilizing Part Confidence Maps and Part Affinity Fields followed by bipartite coordinating and parsing. The identified keypoints are passed to our model where CNN finds examples and LSTM examinations their change after some time. At long last, the model and preparing strategy for framewise expectation and surveying approach on 45 edges (1.5 s) of yield are examined.

10.1 Pose Extraction .

This is the first step of our pipeline and the OpenPose library is used for it. On account of recorded recordings, this progression happens offline, though for continuous forecasts, it happens web based utilizing contribution from the camera to flexibly keypoints to the proposed model. OpenPose is an opensource library for multi-individual keypoint location, which recognizes the human body, hand, and facial keypoints together [49]. The places of 18 keypoints followed by the OpenPose, for example ears, eyes, nose, neck, shoulders, hips, knees, lower legs, elbows, and wrists are shown in Fig. 1. The yield comparing to each casing of a video is gotten in JSON design which contains each body part areas for each individual identified in the picture. The posture extraction was performed at the default goal of OpenPose network for ideal execution. The framework worked at around 3 FPS at these settings. Figure 2 shows the proposed framework design where OpenPose is utilized for keypoint extraction followed by the CNN and LSTM model to anticipate the client's asanas. We utilized recordings with particular subjects for preparing, test, and approval sets with a 60:20:20 split at the video level . After the preprocessing, we get around 8000, 2500, and 2300 casings for preparing testing and approval cases, separately. This strayed from 60:20:20 at the video level because of the variety long of the recordings.

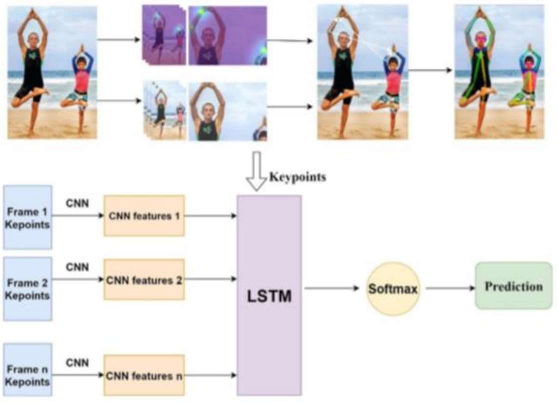


Fig. 2 System architecture: OpenPose followed by CNN and LSTM model

Fig 12. System architecture: openPose followed by CNN and LSTM model

10.2 Model

The profound learning model utilized here is a mix of CNN and LSTM (Fig. 2). CNN is generally utilized for design acknowledgment issues and LSTM is utilized for timeseries errands. In our work, the time-circulated CNN layer is utilized to remove highlights from the 2-D directions of the keypoints acquired in the past advance. The LSTM layer examinations the adjustment in these highlights over the casings, and the likelihood of every Yoga in an edge is given by the Softmax layer. Thresholding is performed on this incentive to distinguish outlines where the client isn't performing Yoga and the impact of surveying on 45 edges has been examined. The model has been modified utilizing Keras Sequential API in Python. The information case has a state of 45 9 18 9 2 which indicates the 45 successive edges with 18 keypoints having X and Y organizes each. Time-distributedCNNlayerwith16filtersofsize3 9 3havingReLU enactment is applied to keypoints of each edge for include extraction. CNNs have a solid capacity to remove spatial highlights which are scale and pivot invariant. The CNN layer can extricate spatial highlights like relative separation and points between the different keypoints in an edge. Cluster standardization is applied to the CNN yield for quicker combination. This is trailed by a dropout layer which haphazardly drops a small amount of the loads forestalling overfitting. The yield from CNN, applied on every one of the 45 casings, is then flattened and passed to LSTM layer with 20 units and unit overlook inclination of 0.5. LSTM is utilized to distinguish transient changes in the highlights extricated by the CNN layer. This use the consecutive idea of video input, and the whole Yoga beginning from its development to holding and delivery is treated as an action. The yield of the LSTM layer comparing to each edge is relaxed circulated completely associated layer with Softmax initiation and six yields. Every one of these six yields gives the likelihood of the comparing Yoga regarding cross-entropy. Thresholding is applied to this yield to recognize when the client isn't performing Yoga. Despite the fact that the model uses LSTM for catching transient relationship, the outcomes are accommodated each edge in the succession and afterward surveyed for the whole grouping of 45 casings. This is additionally expounded in the outcomes area.

10.3 Training

Our assignment is to perceive the client's asanas with appropriate exactness progressively. To begin with, keypoint highlights are removed utilizing OpenPose and recorded the joint area esteems in the JSON file, and afterward CNN and LSTM models are applied for the expectation of asanas. Because of the mix of both, we get the best arrangement of highlights filtered by CNN and long haul information conditions set up utilizing LSTM. The model is aggregated utilizing Keras with Theano backend. The clear cut cross-entropy misfortune work is utilized in view of its appropriateness to gauge the exhibition of the completely associated layer's yield with Softmax initiation. Adam streamlining agent with an underlying learning pace of 0.0001, a beta of 0.9 and no rot are utilized to control the learning rate.

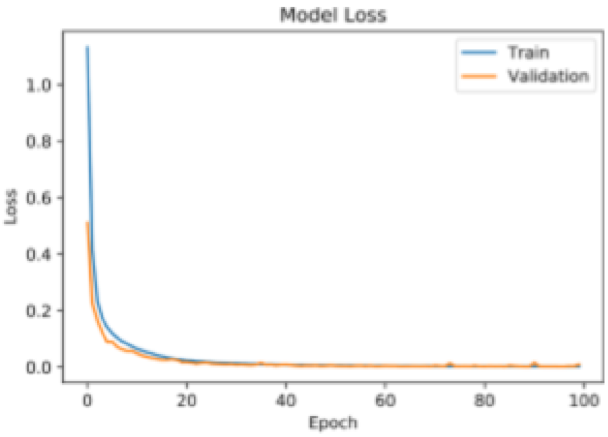
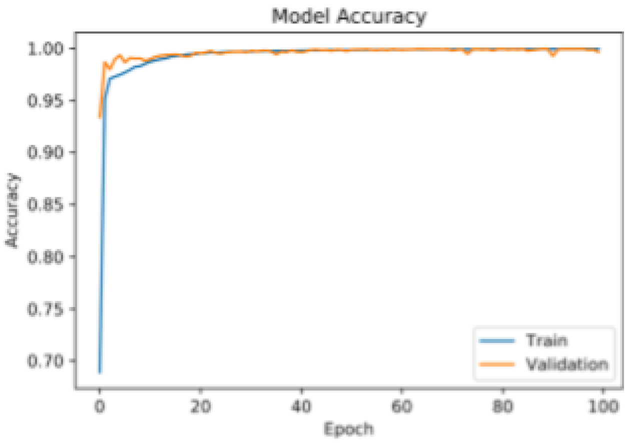


Fig. 13 model loss over the epochs for framewise approach

The model has been prepared for 100 ages on a framework with Intel i7 6700HQ processor, 16GB RAM, and Nvidia GTX 960M GPU. The preparation takes around 22 s for every age which is moderately speedy because of the basic information sources and minimal plan. Figures 3 and 4 show the adjustment in precision and misfortune work, individually, throughout preparing. At first, the preparation and approval exactnesses increment quickly with approval precision remaining over the preparation exactness showing a decent speculation. Afterward, the development is steady, and combination happens after 20 ages. The precision and misfortune way to deal with their asymptotic qualities after 40 ages with minor commotion in the middle. The loads of the best fitting model with most noteworthy approval precision are protected for additional testing. Both, preparing just as approval misfortune have diminished consistently and combined demonstrating a well-fitting model.

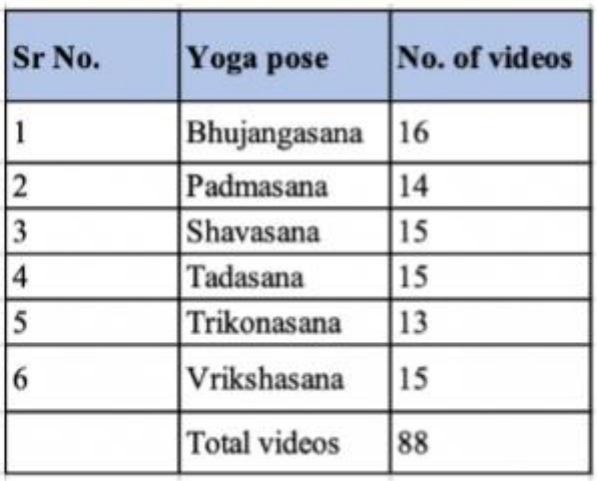


XI. Data Analysis

11.1 Dataset

The dataset utilized for this undertaking is a piece of the Open Source assortment and is freely accessible [50]. This dataset has been made by [2]. It comprises of recordings of 6 yoga presents performed by 15 distinct people (5 females and 10 guys). The 6 yoga presents specifically are – Bhujangasana (Cobra present), Padmasana (Lotus present), Shavasana (Corpse present), Tadasana (Mountain present), Trikonasana (Triangle posture) and Vrikshasana (Tree present). The all out number of recordings is 88 with a term of 1 hour 6 minutes and 5 seconds. The rate at which the recordings have been recorded is 30 FPS (outlines every second). All the recordings have been recorded in an indoor climate at a separation of 4 meters from the camera. All people have performed yoga presents with varieties to help fabricate a dataset which can be utilized to assemble a strong yoga present acknowledgment framework. The normal length of all recordings is around 45 to 60 seconds. Fig. 10 speaks to the quantity of recordings for every yoga present performed by the quantity of people [2].

Table 6. Dataset details



Sr No.	Yoga pose	No. of videos
1	Bhujangasana	16
2	Padmasana	14
3	Shavasana	15
4	Tadasana	15
5	Trikonasana	13
6	Vrikshasana	15
	Total videos	88

The edges from recordings of people performing distinctive yoga presents is appeared in table.

11. Recordings with various subjects have been utilized for the train, test and approval sets.

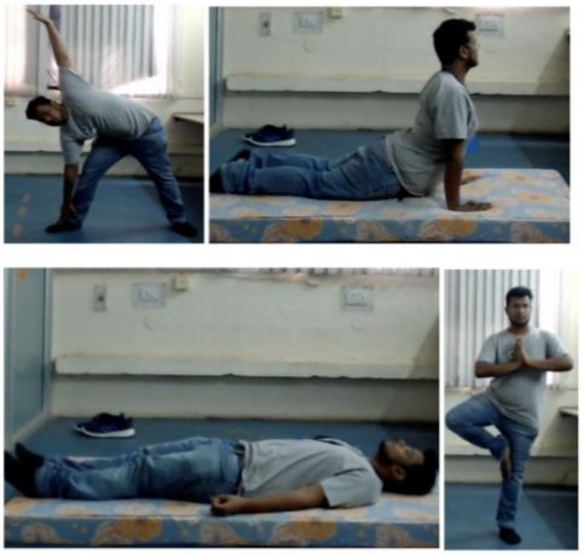


Fig 14. Dataset Frames

11.2 Information Preprocessimg

The initial phase in preprocessing the information is removing keypoints of stances in video outlines utilizing the OpenPose library. For recorded recordings, present extraction is done disconnected while no doubt time, it is done online wherein keypoints distinguished from the contributions to the camera are provided to the model. OpenPose is run on each casing of the video and the comparing yield of each edge is put away in JSON design. This JSON information incorporates the areas of body portions of every individual distinguished in the video outline. Default setting of OpenPose has been utilized for removing present keypoints for ideal execution. Fig. 12 portrays the 18 keypoint positions caught by OpenPose [2].

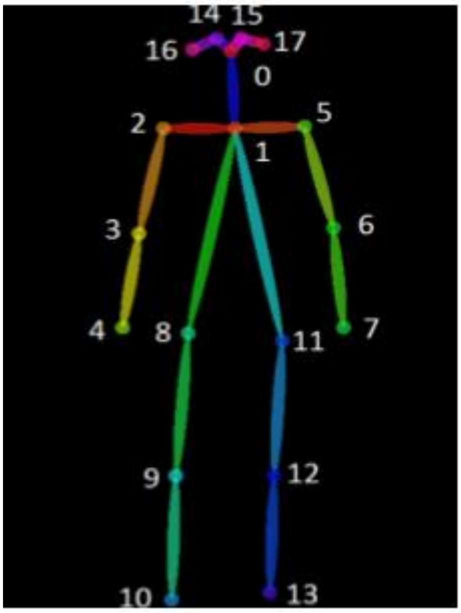


Fig 15. Keypoints identified by OpenPose

The JSON information is gotten and put away in numpy clusters in arrangements of 45 edges which is about 1.5 seconds of the video [2]. 60% of the dataset has been utilized for preparing, 20% for testing what's more, 20% for approval. The preparation information has 7989 successions of 45 casings, each containing the 2D directions of the 18 keypoints caught by OpenPose. The approval information comprises of 2224 such arrangements and the test information contains 2598 groupings. The quantity of casings fluctuated from 60,20,20 split at the video level. This was a direct result of the distinction in span of recordings. Fig. 13 shows the keypoints recognized by OpenPose on each of the 6 yoga presents [2].

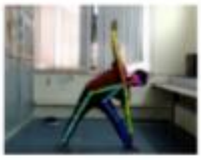
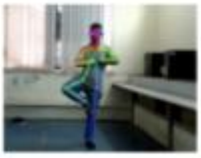
Sr. No.	Yoga asana	Pose
1	Bhujangasana (Cobra pose)	
2	Padmasana (Lotus pose)	
3	Shavasana (Corpse pose)	
4	Tadasana (Mountain pose)	
5	Trikonasana (Triangle pose)	
6	Vrikshasana (Tree pose)	

Fig 16. OpenPose on different pose

XII. Interpretation And Recommendation

12.1 Model Perforamence and Result

Preparing Setup:

The models are manufactured utilizing Python libraries, for example, TensorFlow - Keras (Theano backend), OpenPose, NumPy, Scikit Learn on a framework with NVIDIA Tesla 1080 GPU having 4 GB memory.

1. Backing Vector Machine (SVM)

SVM is an administered AI model that is naturally a two-class classifier. In any case, as most issues include different classes, a multiclass SVM is regularly utilized. A multiclass SVM structures various two class classifiers and separates the classifiers based on the unmistakable name versus the rest (one-versus rest or one-versus all) or between each pair of classes (one-versus one). SVM plays out the grouping by making a hyperplane so that division between classes is as wide as could be expected under the circumstances.

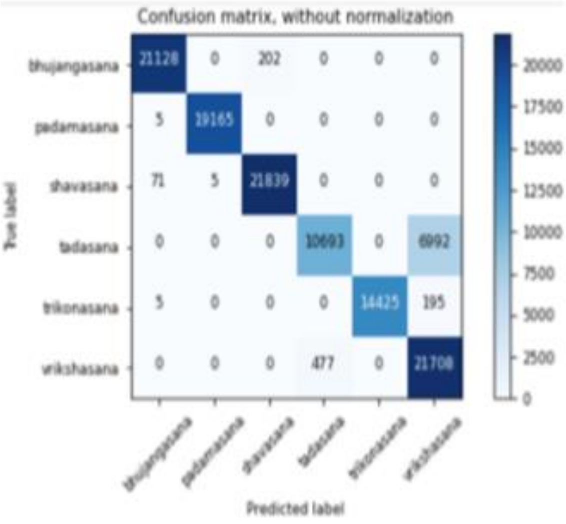
A default SVM has been prepared on the preparation information with the outspread premise work (rbf) portion. Rbf is the default and most mainstream bit which is a gaussian spiral premise work. It gives greater adaptability when contrasted with different parts, direct and polynomial. The estimation of the delicate edge boundary C is 1 and the choice capacity is one-versus rest. The keypoints caught utilizing OpenPose are utilized as highlights to SVM. These 18 keypoints are spoken to by X and Y organizes which makes the absolute number of highlights as 36 (18 * 2). The information is reshaped to make the quantity of tests equivalent.

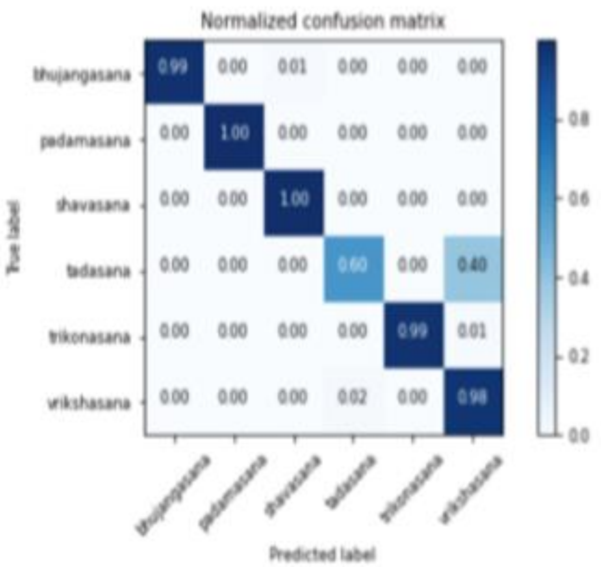
A. Results:

Train precision: 0.9953

Validation exactness: 0.9762

Test precision: 0.9319





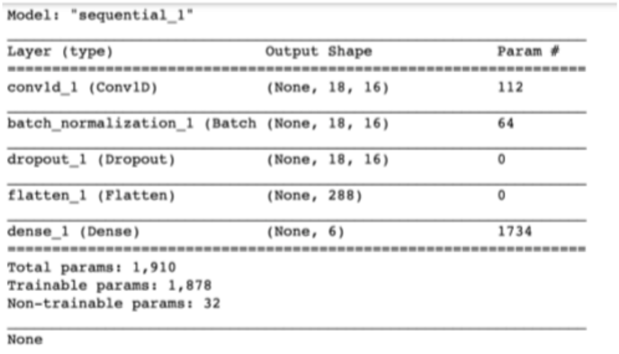
B. Analysis

The preparation precision of the model is pretty high at 0.99. There is a slight abatement in the approval and test precision, yet the outcomes are still acceptable. We can find in the disarray network that most classes are grouped effectively aside from tadasana (mountain present). Out of 17,685 edges for tadasana, 6992 have been misclassified as vrikshasana (tree present) and comparatively there is a few inaccurate arrangement for vrikshasana. This could be a direct result of the closeness in the postures as both of them require a standing position and furthermore the underlying posture arrangement is comparative.

2. Convolutional Neural Network

A one dimensional, one-layer CNN with 16 channels of size 3 x 3 is prepared on the OpenPose keypoints. The info shape is 18 x 2 which connotes the 18 keypoints having X and Y arranges. Bunch standardization is applied to the yield of the CNN layer with the goal that the model merges quicker.

We likewise have a dropout layer that forestalls overfitting by arbitrarily dropping some division of the loads. The actuation work utilized is Rectified Linear Unit (ReLU) which is applied for highlight extraction on keypoints of each edge. The last yield is smoothed prior to being passed to the thick layer with softmax actuation and 6 units where each unit speaks to the probability of a yoga present in cross entropy terms for each of the 6 classes. The model design rundown is appeared in Fig. 14.



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 18, 16)            112
_________________________________________________________________
batch_normalization_1 (Batch (None, 18, 16)            64
_________________________________________________________________
dropout_1 (Dropout)          (None, 18, 16)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 288)               0
_________________________________________________________________
dense_1 (Dense)              (None, 6)                 1734
=================================================================
Total params: 1,910
Trainable params: 1,878
Non-trainable params: 32
_________________________________________________________________
None
```

Fig 17. Model Layers

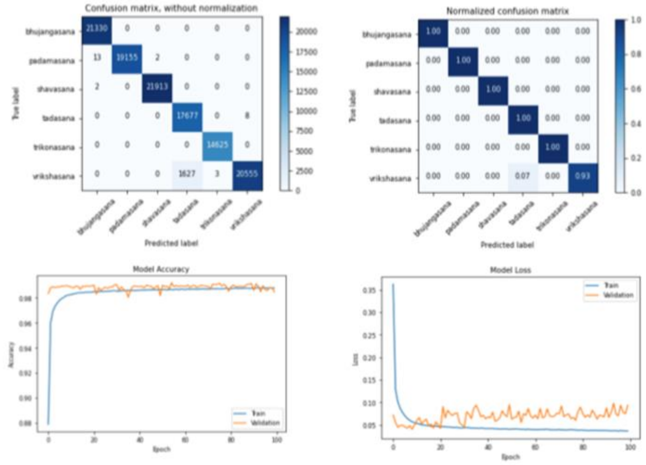
The misfortune work utilized for gathering the model is absolute cross - entropy which is moreover called softmax misfortune. This is utilized as it permits estimating the exhibition of the yield of the thickly associated layer with softmax initiation. This misfortune work is utilized for multi class arrangement, and as we have numerous yoga present classes, it bodes well to utilize downright cross entropy. At long last, to deal with the learning rate, adam enhancer with an underlying learning pace of 0.0001 is utilized. The all out number of ages for which the model is prepared is 100.

A. Results:

Train exactness: 0.9878

Validation precision: 0.9921

Test precision: 0.9858



B. Analysis

The preparation, approval and test precision of the model are nearly the equivalent, around 0.99. The disarray lattice further shows that the model works admirably of grouping all tests accurately, aside from certain examples in vrikshasana which are misclassified as tadasana, prompting 93% precision for vrikshasana. When contrasted with SVM, there are less misclassifications in CNN. Notwithstanding, the model misfortune bend above shows an expansion in the approval misfortune, and a diminishing in the preparation misfortune which shows that there is some overfitting.

3. Convolutional Neural Network + Long Short Term Memory

A profound learning model, CNN and LSTM is utilized [2]. CNN is utilized for investigating outlines furthermore, distinguishing designs, while LSTM makes expectations on the transient information. The CNN layer removes highlights from the keypoints and passes it to the LSTM cells which look at varieties in the traits over various casings. The state of the CNN input is 45 x 18 x 2 which speaks to 45 edges, 18 keypoints in each edge and each keypoint having 2 directions: X and Y. The convolution layers are time disseminated which sends the yield from CNN more than 45 edges (1.5 seconds) of information to the LSTM as a succession. The time dispersed layer is gainful for activities with developments and is subsequently utilized. The CNN yield is smoothed into a one dimensional vector and given as contribution to the LSTM layer which has 20 units, every unit having an overlook inclination of 0.5. The worldly changes in the characteristics distinguished by the CNN are recognized by LSTM which makes a difference influence the consecutive idea of the information video information, in this manner treating the whole yoga present start from its commencement to act changes and delivery like a total movement. The remainder of the CNN engineering is equivalent to show 2. A diagrammatic portrayal of the model is appeared in Fig. 15 [2].

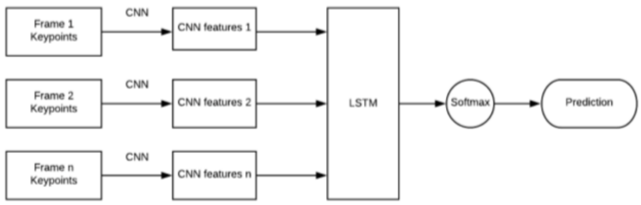
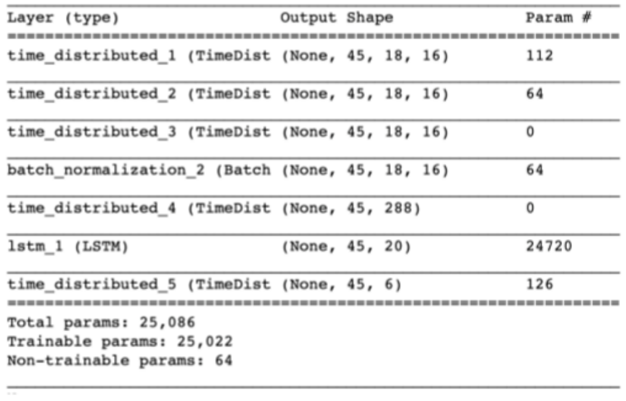


Fig 18. Model Architecture Diagram



```
Layer (type)                 Output Shape              Param #
=================================================================
time_distributed_1 (TimeDist (None, 45, 18, 16)        112
_________________________________________________________________
time_distributed_2 (TimeDist (None, 45, 18, 16)        64
_________________________________________________________________
time_distributed_3 (TimeDist (None, 45, 18, 16)        0
_________________________________________________________________
batch_normalization_2 (Batch (None, 45, 18, 16)        64
_________________________________________________________________
time_distributed_4 (TimeDist (None, 45, 288)           0
_________________________________________________________________
lstm_1 (LSTM)                (None, 45, 20)            24720
_________________________________________________________________
time_distributed_5 (TimeDist (None, 45, 6)             126
=================================================================
Total params: 25,086
Trainable params: 25,022
Non-trainable params: 64
_________________________________________________________________
```

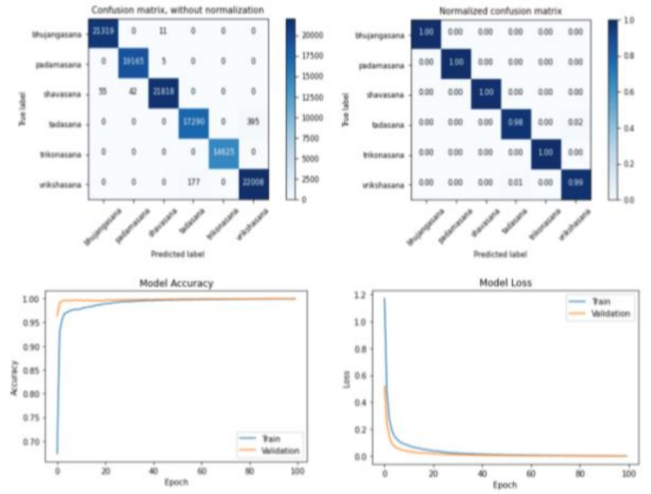
Fig. 16. Model (CNN + LSTM) layers

A. Results:

Train exactness: 0.9992

Validation precision: 0.9987

Test exactness: 0.9938



B. Analysis

The grouping scores are practically near 1 along these lines indicating ideal arrangement for all classes. The slanting in the standardized disarray network is 1.0 for all classes aside from 0.99 for vrikshasana. The quantity of misclassifications for vrikshasana is just 177 which is significantly less as contrasted with the past two models. Additionally, the model precision and model misfortune bend show a decent fit without any changes.

XIII. CONCLUSION

Human posture assessment has been concentrated widely over the previous years. When contrasted with other PC vision issues, human posture assessment is distinctive as it needs to limit and amass human body parts based on an effectively characterized structure of the human body. Use of posture assessment in wellness and sports can help forestall wounds and improve the execution of individuals' exercise. [3] recommends, yoga self-guidance frameworks convey the potential to make yoga famous alongside ensuring it is acted in the correct way. Profound learning techniques are promising a result of the huge exploration being done in this field. The utilization of mixture CNN and LSTM model on OpenPose information apparently is profoundly successful and arranges all the 6 yoga presents impeccably. A fundamental CNN and SVM likewise perform well past our desires.

Execution of SVM demonstrates that ML calculations can likewise be utilized for present assessment or movement acknowledgment issues. Additionally, SVM is a lot lighter and less unpredictable when contrasted with a neural network and requires less preparing time.

XIV. FUTURE WORK

The proposed models right now characterize just 6 yoga asanas. There are various yoga asanas, and subsequently making a posture assessment model that can be effective for all the asanas is a testing issue. The dataset can be extended my adding more yoga presents performed by people in indoor setting as well as open air. The exhibition of the models depends upon the nature of OpenPose present assessment which may not perform well in instances of cover between individuals or cover between body parts. A convenient gadget for self-preparing and constant forecasts can be executed for this framework. This work exhibits movement acknowledgment for reasonable applications. A methodology practically identical to this can be used for present acknowledgment in undertakings for example, sports, reconnaissance, medical services and so forth Multi-individual posture assessment is a totally different issue in itself and has a great deal of degree for research. There are a ton of situations where single individual posture assessment would not get the job done, for instance present assessment in jam-packed situations would have various people which will include following and distinguishing posture of every person. A great deal of factors, for example, foundation, lighting, covering figures and so on which have been talked about before in this overview would additionally make multi-individual posture assessment testing.

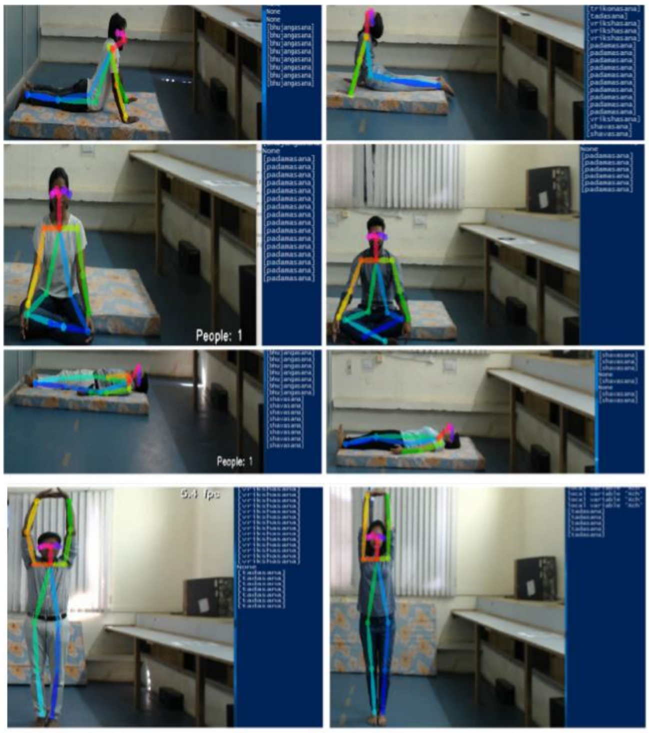


Fig 19. Predictions of asanas in real time (top to bottom row): Bhujangasana (column 2 has the wrong prediction), Padmasana, Shavasana, Trikonasana, and Vrikshasana

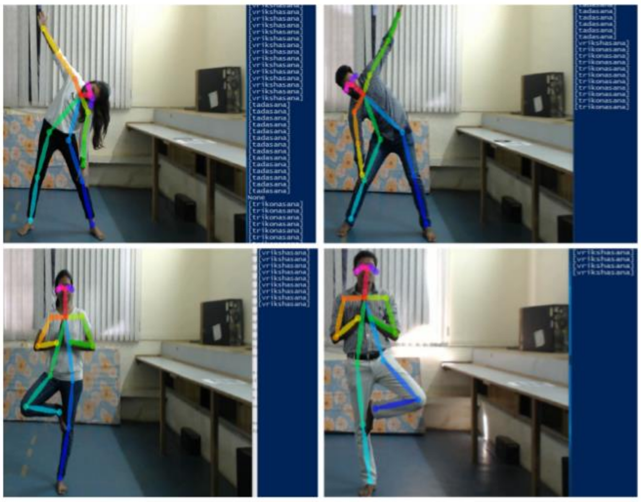
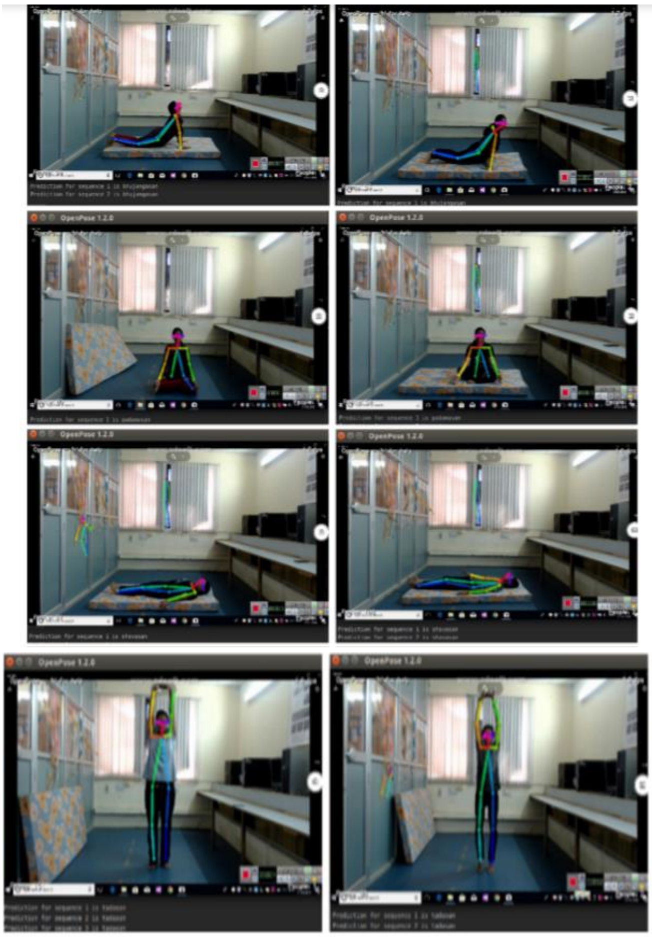
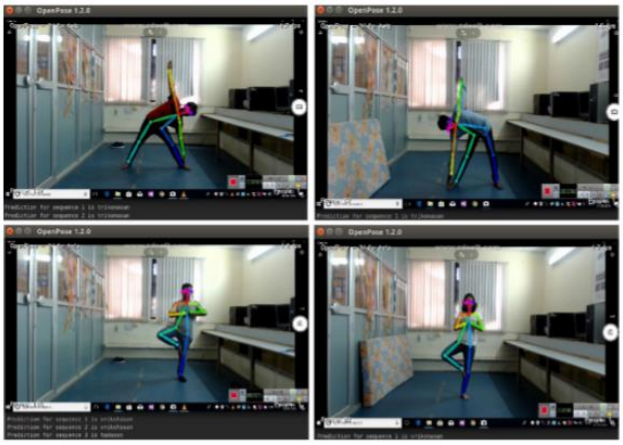


Fig 20. Continued

The methodology of utilizing CNN and LSTM on present information got from OpenPose for Yoga pose identification has been discovered to be exceptionally compelling. The framework perceives the six asanas on recorded recordings just as progressively for 12 people (five guys and seven females). Various people have been utilized for information assortment and ongoing testing. The framework effectively identifies Yoga presents in a video with 99.04% exactness for framewise and 99.38% exactness in the wake of surveying of 45 edges. The framework accomplished 98.92% precision progressively for a bunch of 12 distinct individuals demonstrating its capacity to perform well with new subjects and conditions. It must be noticed that our methodology annihilates the requirement for Kinect or some other specific equipment for Yoga pose identification and can be actualized on contribution from an ordinary RGB camera. In future work, more asanas and a bigger dataset involving both picture and recordings can be incorporated. Additionally, the framework can be executed on a convenient gadget for continuous expectations and self-preparing. This work fills in as an exhibit of action acknowledgment





```
Prediction for sequence 1 is trikonasan
Prediction for sequence 2 is trikonasan
```

```
Prediction for sequence 1 is trikonasan
```

```
Prediction for sequence 1 is vrikshasan
Prediction for sequence 2 is vrikshasan
Prediction for sequence 3 is tadasan
```

```
Prediction for sequence 1 is vrikshasan
```

Fig 20. Predictions of asanas on recorded videos (top to bottom row): Bhujangasana, Padmasana, Shavasana, Tadasana (column 2 has a wrongly predicted sequence), Trikonasana, and Vrikshasana (Column 1 has a wrongly predicted sequence)

XV.REFERENCES

[1]. L. Sigal. “Human pose estimation”, Ency. of Comput. Vision, Springer 2011.

[2]. S. Yadav, A. Singh, A. Gupta, and J. Raheja, “Real-time yoga recognition using deep learning”, Neural Comput. and Appl., May 2019.Online]. Available: https://doi.org/10.1007/s00521-019-04232-7

[3]. U. Rafi, B. Leibe, J.Gall, and I. Kostrikov, “An efficient convolutional network for human pose estimation”, British Mach. Vision Conf., 2016.

[4]. S. Haque, A. Rabby, M. Laboni, N. Neehal, and S. Hossain, “ExNET: deep neural network for exercise pose detection”, Recent Trends in Image Process. and Pattern Recog., 2019.

[5]. M. Islam, H. Mahmud, F. Ashraf, I. Hossain and M. Hasan, "Yoga posture recognition by detecting human joint points in real time using microsoft kinect", IEEE Region 10 Humanit. Tech. Conf., pp. 668-67, 2017.

[6]. S. Patil, A. Pawar, and A. Peshave, “Yoga tutor: visualization and analysis using SURF algorithm”,Proc. IEEE Control Syst. Graduate Research Colloq.,pp. 43-46, 2011.

[7]. W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and H. Zahzah, “Human pose estimation from monocular images: a comprehensive survey”, Sensors, Basel, Switzerland, vol. 16, 2016.

[8]. G. Ning, P. Liu, X. Fan and C. Zhan, “A top-down approach to articulated human pose estimation and tracking”, ECCV Workshops, 2018.

[9]. A. Gupta, T. Chen, F. Chen, and D. Kimber, “Systems and methods for human body pose estimation”, U.S. patent, 7,925,081 B2, 2011.

[10]. H. Sidenbladh, M. Black, and D. Fleet, “Stochastic tracking of 3D human figures using 2D image motion”, Proc 6th European Conf. Computer Vision, 2000.

[11]. A. Agarwal and B. Triggs, “3D human pose from silhouettes by relevance vector regression”, Intl Conf. on Computer Vision & Pattern Recogn.pp.882–888, 2004.

[12]. M. Li, Z. Zhou, J. Li and X. Liu, “Bottom-up pose estimation of multiple person with bounding box constraint”, 24th Intl. Conf. Pattern Recogn.,2018.

[13]. Z. Cao, T. Simon, S. Wei, and Y. Sheikh, “OpenPose: realtime multi-person 2D pose estimation using part affinity fields”, Proc. 30th IEEE Conf. Computer Vision and Pattern Recogn,2017.

[14]. A. Kendall, M. Grimes, R. Cipolla, “PoseNet: a convolutional network for real-time 6DOF camera relocalization”, IEEE Intl. Conf. Computer Vision, 2015.

[15]. S. Kreiss, L. Bertoni, and A. Alahi, “PifPaf: composite fields for human pose estimation”, IEEE Conf. Computer Vision and Pattern Recogn, 2019.

[16]. P. Dar, “AI guardman – a machine learning application that uses pose estimation to detect shoplifters”.Online]. Available:

https://www.analyticsvidhya.com/blog/2018/06/ai-guardman-machine-learning application-estimates-poses-detect-shoplifters/

[17]. D. Mehta, O. Sotnychenko, F. Mueller and W. Xu, "XNect: real-time multi-person 3D human pose estimation with a single RGB camera", ECCV, 2019.

[18]. A. Lai, B. Reddy and B. Vlijmen, "Yog.ai: deep learning for yoga".Online]. Available: http://cs230.stanford.edu/projects_winter_2019/reports/15813480.pdf

[19]. M. Dantone, J. Gall, C. Leistner, "Human pose estimation using body parts dependent joint regressors", Proc. IEEE Conf. Computer Vision Pattern Recogn., 2013.

[20]. A. Mohanty, A. Ahmed, T. Goswami, "Robust pose recognition using deep learning", Adv. in Intelligent Syst. and Comput, Singapore, pp 93-105, 2017.

[21]. P. Szczuko, "Deep neural networks for human pose estimation from a very low resolution depth image", Multimedia Tools and Appl, 2019.

[22]. M. Chen, M. Low, "Recurrent human pose estimation",Online]. Available: https://web.stanford.edu/class/cs231a/prev_projects_2016/final%20(1).pdf

[23]. K. Pothanaicker, "Human action recognition using CNN and LSTM-RNN with attention model", Intl Journal of Innovative Tech. and Exploring Eng, 2019.

[24]. N. Nordsborg, H. Espinosa, "Estimating energy expenditure during front crawl swimming using accelerometrics", Procedia Eng., 2014.

[25]. P. Pai, L. Changliao, K. Lin, "Analyzing basketball games by support vector machines with decision tree model", Neural Comput. Appl., 2017.

[26]. S. Patil, A. Pawar, A. Peshave, "Yoga tutor: visualization and analysis using SURF algorithm", Proc. IEEE Control Syst. Grad. Research Colloquium, 2011.

[27]. W. Wu, W. Yin, F. Guo, "Learning and self-instruction expert system for yoga", Proc. Intl. Work Intelligent Syst. Appl, 2010.

[28]. E. Trejo, P. Yuan, "Recognition of yoga poses through an interactive system with kinect device", Intl. Conf. Robotics and Automation Science, 2018.

[29]. H. Chen, Y. He, C. Chou, "Computer assisted self-training system for sports exercise using kinetics", IEEE Intl. Conf. Multimedia and Expo Work, 2013.

[30]. DatasetOnline]. Available: https://archive.org/details/YogaVidCollected.

[31]. Y. Shavit, R. Ferens, "Introduction to camera pose estimation with deep learning",Online]. Available: https://arxiv.org/pdf/1907.05272.pdf.

[32]. Gao Z, Zhang H, Liu AA et al (2016) Human action recognition on depth dataset. Neural Comput Appl 27:2047–2054. https://doi. org/10.1007/s00521-015-2002-0

[33]. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28:976–990. https://doi.org/10.1016/j. imavis.2009.11.014

[34]. Weinland D, Ronfard R, Boyer E (2011) A survey of visionbased methods for action representation, segmentation and recognition. Comput Vis Image Underst 115:224–241. https://doi. org/10.1016/j.cviu.2010.10.002

[35]. Halliwell E, Dawson K, Burkey S (2019) A randomized experimental evaluation of a yoga-based body image intervention. Body Image 28:119–127. https://doi.org/10.1016/j.bodyim.2018.12. 005

[36]. Patil S, Pawar A, Peshave A et al (2011) Yoga tutor: visualization and analysis using SURF algorithm. In: Proceedings of 2011 IEEE control system graduate research colloquium, ICSGRC 2011, pp 43–46

[37]. Chen HT, He YZ, Hsu CC et al (2014) Yoga posture recognition for self-training. In: Lecture notes in computer science (including

subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), pp 496–505

[38]. Schure MB, Christopher J, Christopher S (2008) Mind–body medicine and the art of self-care: teaching mindfulness to counseling students through yoga, meditation, and qigong. J Couns Dev. https://doi.org/10.1002/j.1556-6678.2008.tb00625.x

[39]. Chen HT, He YZ, Hsu CC (2018) Computer-assisted yoga training system. Multimed Tools Appl 77:23969–23991. https://doi.org/10.1007/s11042-018-5721-2

[40]. Maanijou R, Mirroshandel SA (2019) Introducing an expert system for prediction of soccer player ranking using ensemble learning. Neural Comput Appl. https://doi.org/10.1007/s00521019-04036-9

[41]. Nordsborg NB, Espinosa HG, Thiel DV (2014) Estimating energy expenditure during front crawl swimming using accelerometers. Procedia Eng 72:132–137. https://doi.org/10.1016/j.proeng.2014. 06.024

[42]. Connaghan D, Kelly P, O'Connor NE et al (2011) Multi-sensor classification of tennis strokes. Proc IEEE Sens. https://doi.org/10.1109/icsens.2011.6127084

[43]. Shan CZ, Su E, Ming L (2015) Investigation of upper limb movement during badminton smash. In: 2015 10th Asian Control conference, pp 1–6. https://doi.org/10.1109/ascc.2015.7244605 .

Authors

Deepak Kumar
Department of Information Technology, Research Scholar, Amity University, Jharkhand, India

Anurag Sinha
Department of Information Technology, Research Scholar, Amity University, Jharkhand, India

Cite this article as :

Deepak Kumar, Anurag Sinha , "Yoga Pose Detection and Classification Using Deep Learning ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 6, pp. 160-184, November-December 2020.
Available at
doi : https://doi.org/10.32628/CSEIT206623
Journal URL : http://ijsrcseit.com/CSEIT206623