ISSN:3007-3391

Journal of Computers and Intelligent Systems
Vol. 3, No. 3, 2025, pages 218 - 233
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

Multi-Scale Human Pose Estimation Using Morphological Segmentation and Deep Learning

Anam Naveed[1], Syed M. Adnan[1], Wakeel Ahmad[1] , Mamoona Sadia[1]

[1]Department of Computer Science University of Engineering and Technology, Taxila.

ARTICLE INFO

Article History:

Received:	June	25, 2025
Revised:	August	28, 2025
Accepted:	August	30, 2025
Available Online:	August	30, 2025

Keywords:

2D human pose estimation
Morphological segmentation
Weber's law
VGG16
CNN

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

The intersection of computer vision, computer graphics, and machine learning leads to human modeling and pose estimation. Human pose estimation has been and continues to be a challenging issue in computer vision because of occlusions, differences in sizes of bodies, and intricate joint movements. Even with recent breakthroughs in deep learning, correctly identifying salient events in real-world settings remains a major challenge. To solve these problems, we introduce a new approach to accurate human pose estimation that combines morphological segmentation with deep learning. Morphological operations help segment the input images, and Convolutional Neural Network (CNN) architecture like VGG16 is utilized to extract significant features from the segmented images, which are then classified using classifiers. The model, which is proposed, is trained on two publicly shared datasets, MPII and LSP, to capture diverse human poses with varying conditions and scales. We emphasize the success of our approach in attaining sophisticated results in human pose estimation tasks by engaging in extensive testing and evaluation. Our method effectively deals with occlusions and intricate poses along with accurately detecting key points. We also highlight the model's interpretability and generalizability, presenting its strength in numerous real-life scenarios.

.

1. Introduction

Human Pose Estimation (HPE) represents a foundational challenge in computer vision, focused on the precise localization of human anatomical keypoints—such as joints and limbs—within images or video sequences [1] . This capability is a crucial required for more advanced activities, including human action recognition, gesture-based human-computer interaction, and advanced behavioral analysis [2] . But there is a crucial difference: whereas pose provides required spatial form, it does not of itself encode the temporal dynamics needed for grasping motion and purpose [3].

The primary objective of 2D HPE is to locate these keypoints and align them onto a standard set of body landmarks [4]. This is beyond basic object detection, requiring not only the detection of individual parts but

also the intricate modeling of their spatial relationships, which are highly variable across poses. Given that 3D HPE only gives a more abstract spatial representation, but one that is very much dependent on good 2D pose estimation as the input of choice, meaning 2D estimation progress translates directly to the 3D realm[5, 6].

Even though it is useful in diverse applications such as security monitoring, sports analytics, and telehealth monitoring, 2D HPE is still full of challenges. Recurring challenges are occlusions, where major body parts are obstructed; Various environmental conditions like low illumination and low contrast; and the natural intricateness of crowded backgrounds[7] . These elements require a trade-off between computational efficiency and estimation accuracy that contemporary research continuously aims to maximize [8] .

Responding directly to these challenges, this paper proposes a new multi-stage framework to improve the robustness and accuracy of 2D human pose estimation. Our approach combines state-of-the-art image pre-processing, using morphological segmentation to clean human silhouettes and benefiting from principles borrowed from Weber's Law for perceptually salient feature detection. A VGG16 convolutional neural network structure is the backbone of our feature extraction pipeline with a deep hierarchy of visual features. Their discriminative ability is then exhaustively tested with an arsenal of machine learning algorithms. By combining these components, our proposed system is expected to provide robust performance with difficult real-world situations under occlusion, varying illumination, and complicated environment.

2. Related Work

Human pose estimation has experienced various methodological stages, beginning with classical computer vision methods before evolving to current deep learning methods. Initial research works were generally divided into two classes. The first class consisted of appearance-based methods that defined body parts using hand-designed feature descriptors. An example is well-known systems like Poselets, which were feature-based and comprised Histogram of Oriented Gradients (HOG) [9, 10].

The second approach utilized structured models, which focused on body part spatial relations. One landmark work in this field is the Pictorial Structures Model [11],which represented the body as a stretchy constellation of parts connected by constraints. This was generalized subsequently by models that used mixtures of parts to explain greater variability of articulation [1]. Though basic, these non-deep learning methods were generally confined to laboratory settings and did not fare well in complicated, multi-individual situations.

Deep learning created an unprecedented revolution with the advent of Convolutional Neural Networks (CNNs) as the most suitable feature extraction backbone due to their ability to learn hierarchical representation from data directly. Extensively examined, e.g., in the work of Munea et al. [8], the architectural finesse, training practices, and typical performance of those CNN-based systems have been reported. In recent years, transformer-based models have been borrowed from natural language processing and have shown strong performance in modeling long-range spatial relations among keypoints, again advancing the state-of-the-art on both 2D and 3D HPE. [12].

Much of the current work targets the key problem of occlusion and prediction refinement. Instead of assuming that an initial pose prediction is final, more recent methods use special refinement modules. For example, PoseFix proposed a generic post-processing network that corrects errors in the output of any upstream model by using contextual image features. Furthermore, other works have merged Graph Convolutional Networks (GCNs) to explicitly model the skeletal structure of the human body, using these learned earlier to infer credible positions for occluded or vague joints[2].

The processing of video sequences presents the dimension of temporal data, which is leveraged to improve prediction stability and accuracy. Although image-based models like the VGG-16 architecture applied to static frames can learn robust feature representations [13], temporal models employ series of frames to sort out ambiguities. Recurrent Neural Networks (RNNs) and other sequence-to-sequence models have been productively employed to spread information across time, smoothing predictions and improving robustness in video-based pose estimation.

The field has also been driven forward by the improvement and extensive adoption of efficient, open-source pose estimation libraries. OpenPose initiated real-time multi-person estimation using Part Affinity Fields for grouping keypoints. It was then followed by subsequent frameworks such as PoseNet, which scale lighter models developed on top of MobileNet and ResNet, and Google MediaPipe Pose, which was optimized to work in real time on edge and mobile devices. These libraries have been central to normalizing access and standardizing evaluation to state-of-the-art models [14].

Finally, the developments in HPE are documented in large-scale comparative and review studies that benchmark advancements against large-scale datasets like COCO and MPII. These studies effectively measure the evolution of efficiency and accuracy, providing a clear indication of the path of the field. This has enabled the deployment of HPE over a wide variety of applications, ranging from very accurate activity recognition in smart domain to hybrid rule-based and deep learning for healthcare monitoring like fall detection [15]. Table 1 presents these comparative insights with the path of HPE from old-fashioned to contemporary methods.

Table 1: Comparison of Related Work

Ref	Years	Technique	Dataset	Accuracy%
[16]	2020	SVM	MSR3D	72.50
[1]	2021	Deep learning	Yoga-82	94.91
[3]	2022	CNN	COCO	69.70
[17]	2022	GCN	LSP,FLIC, MPII, COCO	69.70
[8]	2023	CNNs, RNNs	LSP, FLIC, MPII, COCO.	73.40
[13]	2023	CNN, RNN, LSTM, YOLOv5, Open Pose, ResNet32 RCNN,Resnet50.	MPII, COCO, human3.6, MCFD, and URFD	75.00
[18]	2023	SVM, KNN	Human 3.6M, MPI-INF-3DHP	79.97
[19]	2024	HRNet, CBAM	COCO 2017	67.10
[20]	2024	PRHP	COCO, MPII	74.00
[21]	2025	Pose Scoring Model	COCO	78.10

3. Used Approach

3.1. Materials and Methods

The solution strategy is shown in Figure 1, which is a general schema for 2D human pose estimation through a four-step pipeline. Image capture is the starting point where input data are obtained from different sources. Morphological segmentation followed by improving the human silhouettes utilizing advanced image processing

techniques for feature enhancement is next. The essential idea within our approach employs a deep model (VGG16-based architecture) to extract hierarchical features and spatial relations between body joints. For identifying posture keypoints correctly and validating, the classification process ultimately employs a collection of machine learning models.

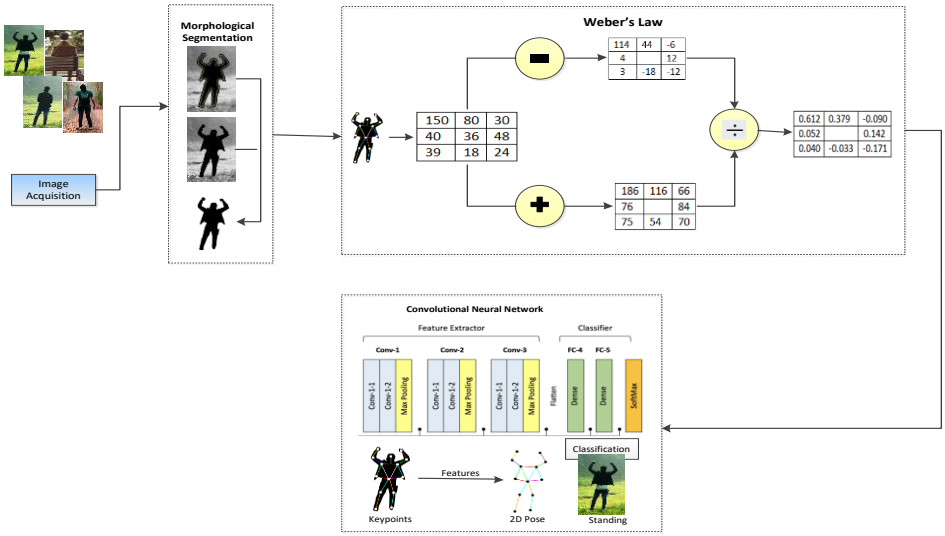


Figure 1: Propose model for Human pose estimation

Image Acquisition

To ensure quality input for our human pose estimation research, extreme caution in image capture was the first step. The data acquisition method of choice has a critical impact on the performance and applicability of our proposed approach. In this work, we intentionally amassed a comprehensive dataset by leveraging two readily available sources: the MPII dataset and the LSP. This combination was selected with purpose to address our varied needs for our research objectives. The MPII dataset is an extensive one with over 25,000 images, all with exquisite multi-person pose annotations. The images capture a wide range of human activity within everyday environments. Assuming joint key point annotations have relevance, providing detailed dimensional data for the generation of robust pose estimation models. Apart from MPII, we have also used the LSP dataset. This is a dense human pose estimation data set consisting of 2,000 images and a corresponding visualized image with human joints labeled distinctly. The LSP dataset enhances our training data through the provision of an extensive set of variable human poses with a focus on those which are found to occur in sports settings. To help ensure consistency and uniformity of our data, we established strict preprocessing protocols. This included resizing, normalization, and other augmentation techniques like flipping and rotation. These are needed to improve our model's ability to generalize from unknown data. The strategic combination of the MPII and LSP datasets significantly improves the diversity of our training set. This variation enriches our model's ability to correctly identify intricate human poses in a wide range of real-world situations, thereby providing robust support to our new approach. The behavior of the model under different circumstances will be demonstrated graphically in Figure 2, and Table 2 will provide a detailed overview of the datasets employed.

Figure 2: Sample images from the MPII dataset [20]

Figure 3: Sample images from the LSP dataset [8]

S#	Category	Total Number of Images
1	*MPII*	*25000*
2	*LSP*	*2,000*
Total		*27,000*

Table 2: Data description of datasets

Morphological Segmentation:

Morphological segmentation is a crucial step of the pre-processing pipeline in 2D Human Pose Estimation to improve the quality of input images and improve key point localization considerably. The method elegantly alters the shapes and structures of images by a series of image processing processes. These techniques are extremely useful to resolve general issues like distracting backgrounds or occlusions, which are commonly found in real-world datasets like MPII and LSP. Our technique uses morphological segmentation to highlight the silhouette of the human body and remove unnecessary background details from gray-scale or binary images obtained from the raw data. This stage is instrumental in directing the focus of the model towards significant body joints by escalating silhouettes and outlining the human body clearly against complex backgrounds. Morphological segmentation gives a more precise input to the pose estimation network, which is critical due to the action specific poses in the MPII dataset and the diverse sports related poses in the LSP dataset that tend to include many individuals and partial occlusions. This eventually results in better joint localization and more robust generalization across a broader range of positions and environments. The combination of morphological segmentation greatly increases the quality and robustness of the overall pose estimation pipeline by giving cleaner, noise-free input representations.

Mathematically dilation of an image X by a structuring element Y is denoted as $X \oplus Y$ illustrated in equation (1).

$$X \oplus Y = \left\{ z \middle| \left(\hat{y} \right)_z \cap X \neq \phi \right\} \quad \text{..... (1)}$$

Mathematically dilation of an image X by a structuring element Y is denoted as $X \ominus Y$ illustrated in equation (2).

$$X \Theta Y = \left\{ z \middle| \left(\hat{y} \right)_z \cap x^c = \phi \right\} \quad \text{..... (2)}$$

Weber's Law:

For better explanation of the implementation, upon processing by morphological segmentation to separate the region of interest, we used Weber's Law to emphasize perceptually relevant pixel differences. According to Weber's Law, an increase in stimulus is perceived as a function of its ratio to the original intensity and not the absolute change.

Weber's Law is commonly expressed as in equation (3):

$$\frac{\Delta I}{I} = k \quad \text{..... (3)}$$

Where ΔI is the intensity change, I is the background intensity, and k is a constant. Here, in our method, this concept was employed to modulate the contrast representation of the segmented images. Small intensity variations in low-contrast areas were enhanced, whereas variations in high-intensity areas were comparatively dampened. This was to make small but significant variations in bone or joint structures more discernible and recognizable for further processing.

Weber's Law–augmented images were subsequently passed to VGG16 for deep feature extraction. Morphological segmentation thus supplied clean input, Weber's Law highlighted perceptually significant intensity information, and VGG16 yielded strong and discriminative features for classification.

VGG 16 Models:

Our method utilizes the VGG16 architecture as a key component for feature extraction in our framework for estimating human posture following the preliminary image pre-processing. The renowned deep convolutional neural network (CNN) VGG16 is known for its simple but efficient method of extracting hierarchical features from visual input. Its structure, designed by the University of Oxford's Visual Geometry Group, is defined by the repeated use of small 3x3 convolutional filters throughout the network. The VGG16 architecture consists of 16 layers with melodious parameters, structured into series of convolutional blocks succeeded by max-pooling processes, ending with multiple fully connected layers. Usually, each convolutional block consists of two or three convolutional layers that use 3×3 filters, followed by a 2×2 max-pooling layer. This step of pooling effectively diminishes spatial dimensions without losing critical characteristics. Depth of the network, obtained by piling up these tiny filters, allows the extraction of increasingly more complex and abstract features in its lower levels.

Fundamental Mathematical Operations:

The elemental operation employed in VGG16 is convolution, as formulated in Equation (4) .For an input feature map F and a convolutional kernel K, the resulting output feature map G at a specific position (x, y) is calculated as:

$$G(x,y) = \sum_{i=0}^{H_k-1} \sum_{J=0}^{W_k-1} F(x+i, y+j).K(i,j) \quad \text{..... (4)}$$

Here, HK and WK represent the height and width of the convolutional kernel K. This is done throughout input feature map, with padding and strides managing the spatial extents of the output. VGG16 predominantly uses 3×3 kernels with a stride of 1 and padding to preserve spatial dimensions before pooling.

The equation (5) represent the max-pooling operation, typically applied after a series of convolutional layers, down samples the feature maps by selecting the maximum value within a defined window (e.g., a 2×2 region):

$$G_{pooled}(x,y) = \max_{(i,j)\in window} F(x.stride+i, y.stride+j) \quad \text{.....(5)}$$

These operations collectively enable VGG16 to efficiently extract a hierarchical set of features, providing a strong basis for precise human pose estimation. The architecture of the VGG16 model is described in Figure 3.

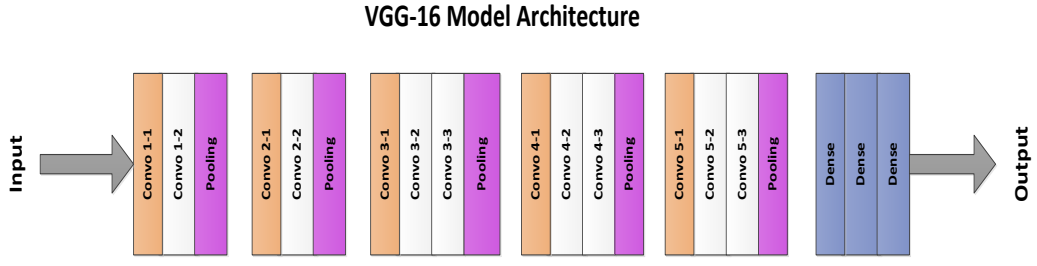


Figure 3: Architecture model of VGG 16

Classification:

Several classification methods were used to assess the effectiveness of the derived features utilizing Weber's law and VGG16. Due to their shown effectiveness in identifying patterns in high-dimensional feature data, these classifiers were selected. The classifiers used in this research are ensemble-based methods, K-Nearest Neighbors (KNN) with different distance measures, Support Vector Machines (SVM) with different kernels, Linear Discriminant Analysis (LDA) Ensemble boosted tree, and Quadratic Discriminant Analysis (QDA).

Classification for the MPII Dataset

The MPII dataset features complex human poses showcasing different angles, obstructions, and cluttered backgrounds. In order to address these challenges, various classifiers are assessed:

Linear Discriminant Analysis (LDA) – A dimension reduction technique that maps feature into a space to increase inter-class discrimination at the expense of intra-class variability. It assumes Gaussian-distributed data and is computationally efficient for pose classification.

Quadratic Discriminant Analysis (QDA) – A variant of LDA called quadratic discriminant analysis (QDA) allows for more flexible decision limits by reducing the need for uniform class covariance's. This makes it easier to distinguish between identical postures with minor changes.

Quadratic SVM – A quadratic kernel-based support vector machine that may depict non-linear relationships between postures attributes. It works particularly well in scenarios with moderately non-linear positions.

Cubic SVM – Improves classification for highly articulated roles by using a third-degree polynomial kernel to detect more complex decision boundaries.

Fine Gaussian SVM – A kernel-driven SVM with a properly calibrated Gaussian (RBF) kernel that offers high classification accuracy for poses with significant intra-class variance.

Medium K-Nearest Neighbors (KNN) – A distance-based classifier that groups postures based on the closest neighbours' majority vote in the feature space. Bias and variance in matching postures are balanced in the "medium" setting.

Cosine KNN – Improves robustness to changes in the magnitude of keypoint positions by evaluating the angular similarity between pose feature vectors rather than Euclidean distance.

Ensemble Subspace KNN – Enhances generalization for varied poses and reduces over fitting by integrating multiple KNN classifiers trained on random feature subsets.

Ensemble Subspace Discriminant – A collection of discriminant classifiers (LDA/QDA) developed on various feature subsets, improving the stability of pose classification.

LSP Dataset Classification

The majority of the athletic poses in the LSP dataset include a lot of articulation and dynamic motion. The classifiers listed below are used:

Ensemble Boosted Trees (such as AdaBoost and Gradient Boosting) – Boosting techniques train sequential weak learners (decision trees) to correct misclassifications, making them useful for detailed pose differentiation. They handle occlusions and challenging poses well due to their iterative error correction mechanism.

K-Nearest Neighbors (KNN) – A simple yet effective classifier that matches position attributes based on proximity, useful for identifying similar stances in sports settings.

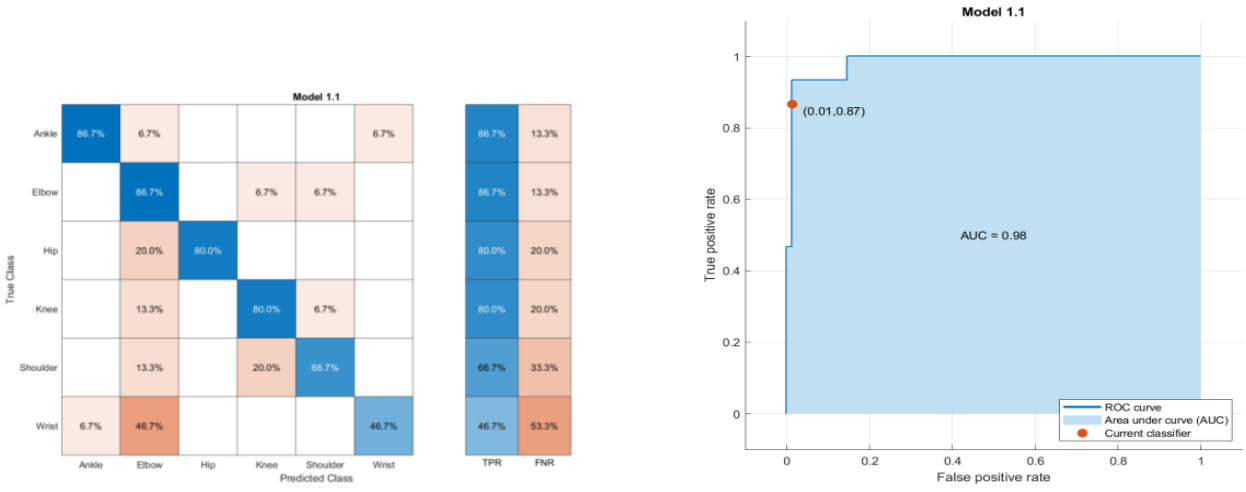
Linear SVM – A fast, easily comprehensible classifier that works well when pose features can be divided linearly, providing a standard against which others can be measured.

Quadratic SVM – Combines a quadratic kernel with linear SVM to improve performance for non-linear pose distributions that are commonly encountered in sports.

3.2. Results and discussion

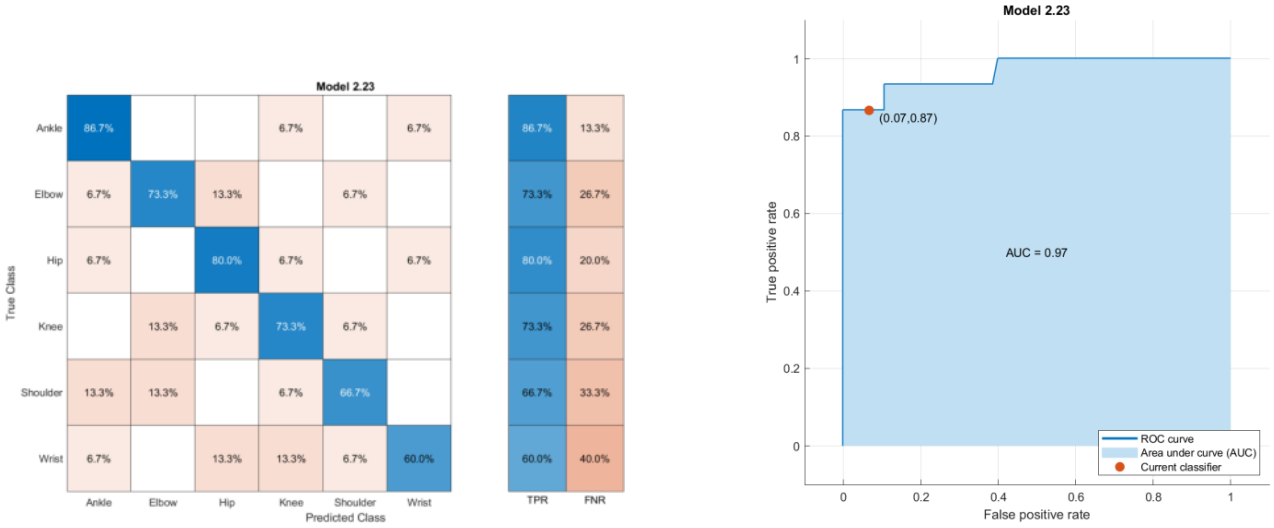
The efficacy of the proposed human pose estimation technique was validated with two diverse and widely used datasets: MPII and LSP. Both data sets pose unique challenges regarding pose variation, occlusion, and background noise. The method used morphological segmentation to separate human figures from the rest of the scene, followed by Weber's Law-based contrast enhancement, and deep feature extraction using the VGG16 convolutional neural network. The extracted features were then used for classification using different machine learning classifiers.

Several classifiers were used to evaluate classification accuracy on the MPII dataset. Linear Discriminant classifier achieved the highest accuracy of 74.4% as shown in figure 4, suggesting effective separation of pose features in a linear model. Quadratic Discriminant classifier achieved 73.2%, which means that non-linear decision boundaries may have affected performance slightly, possibly due to too high a feature dimensionality. Ensemble Subspace KNN recorded promising results at 73.3% as seen in figure 5, depicting the ability of ensemble techniques to handle complex feature patterns.



(a) (b)

Figure 4: Performance of LDA (a) Confusion Matrix (b) ROC Curve



(a) (b)

Figure 5: Performance of Ensemble Subspace KNN (a) Confusion Matrix (b) ROC Curve

Other classifiers such as Quadratic SVM (72.9%), Cubic SVM (72.2%), and Medium KNN (71.1%) gave similar but slightly lower accuracies, while Fine Gaussian SVM and Cosine KNN gave 70.0%, suggesting that fine-tuning the kernel or distance metric will not better represent the MPII pose variations.Ensemble Subspace Discriminant, though based on linear methods, achieved 70.9%, supporting the observed trend that linear classifiers performed well with the feature set given, as evidenced from Table 3.

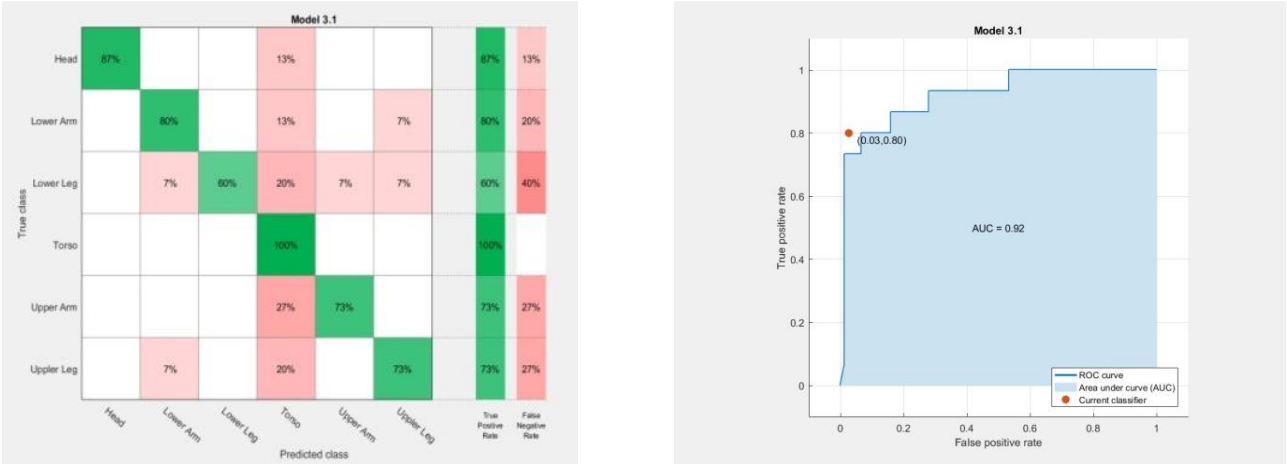
Table 3: Summary of the classification report for the MPII classifier.

Classifier	Classes	True Positive Rate (%)	False Negative Rate (%)	Positive Predictive Value (%)	False discovery rate (%)	Accuracy (%)
Linear Discriminant	Ankle	86.7	13.3	92.9	7.1	**74.4**
	Elbow	86.7	13.3	46.4	53.6	
	Hip	80.0	20.0	100.0		
	Knee	80.0	20.0	75.0	25.0	
	Shoulder	66.7	33.3	83.3	16.7	
	Wrist	46.7	53.3	87.5	12.5	
Quadratic Discrimi	Ankle	93.3	6.7	77.8	22.2	73.2
	Elbow	73.3	26.7	45.8	54.2	

	Hip	60.0	40.0	100.0		
	Knee	86.7	13.3	92.9	7.1	
	Shoulder	66.7	33.3	90.9	9.1	
	Wrist	60.0	40.0	64.3	35.7	
Quadratic SVM	Ankle	86.7	13.3	92.9	7.1	72.9
	Elbow	66.7	33.3	83.3	16.7	
	Hip	80.0	20.0	80.0	20.0	
	Knee	80.0	20.0	41.1	58.6	
	Shoulder	73.3	26.7	91.7	8.3	
	Wrist	53.3	46.7	100.0		
Cubic SVM	Ankle	86.7	13.3	100.0		72.2
	Elbow	66.7	33.3	76.6	23.1	
	Hip	80.0	20.0	85.7	14.3	
	Knee	93.3	6.7	43.8	56.2	
	Shoulder	53.3	46.7	88.9	11.1	
	Wrist	53.3	46.7	88.9	11.1	
Fine Gaussian SVM	Ankle	80.0	20.0	100.0		70.0
	Elbow	66.7	33.3	100.0		
	Hip	80.0	20.0	100.0		
	Knee	93.3	6.7	38.9	61.1	
	Shoulder	33.3	66.7	83.3	16.7	
	Wrist	66.7	33.3	71.4	28.6	
Medium KNN	Ankle	86.7	13.3	65.0	35.0	71.1
	Elbow	66.7	33.3	100.0		
	Hip	86.7	13.3	68.4	31.6	
	Knee	73.3	26.7	64.7	35.3	

	Shoulder	66.7	33.8	62.5	37.5	
	Wrist	46.7	53.3	87.5	12.5	
Cosine KNN	Ankle	86.7	13.3	56.5	43.5	70.0
	Elbow	60.0	40.0	69.2	30.8	
	Hip	86.7	13.3	68.4	31.6	
	Knee	73.3	26.7	68.8	31.2	
	Shoulder	60.0	40.0	100.0		
	Wrist	53.3	46.7	80.0	20.0	
Ensemble Subspace KNN	Ankle	86.7	13.3	72.2	27.8	73.3
	Elbow	73.3	26.7	73.3	26.7	
	Hip	80.0	20.0	70.6	29.4	
	Knee	73.3	26.7	68.8	31.2	
	Shoulder	66.7	33.3	76.9	23.1	
	Wrist	60.0	40.0	81.8	18.2	
Ensemble Subspace Discriminant	Ankle	80.0	20.0	60.0	40.0	70.9
	Elbow	66.7	33.3	90.9	9.1	
	Hip	80.0	20.0	80.0	20.0	
	Knee	86.7	13.3	54.2	45.8	
	Shoulder	66.7	33.3	76.9	23.1	
	Wrist	46.7	53.3	100.0		

In contrast, Table 4 displays the findings from the LSP dataset, indicating relatively better classification outcomes, probably a result of differences in image resolution, uniformity of annotations, and variety of poses. Figure 6 demonstrates that the Linear SVM achieved the best performance with an accuracy of 78.9%, closely followed by the Quadratic SVM at 78.6%, highlighting the effectiveness of support vector machines in detecting complex pose-related patterns in this dataset. The Ensemble Boosted Tree classifier proved effective too, reaching 75.6% by applying the ensemble method to improve classification accuracy. The KNN classifier achieved 74.4%, showing good performance with distance metrics, although it was slightly less effective than SVM approaches.



(a) (b)

Figure 6: (a) Confusion Matrix of Linear SVM, (b) ROC curve of Linear SVM

Performance of the classifier on the MPII and LSP data was examined by means of confusion matrices and ROC curves. The confusion matrices indicate that the model is making reasonable joint-level prediction in both data sets. The respective ROC curves show that the proposed approach is able to correctly differentiate correct joint positions from incorrect ones effectively, and thus present quantitative evidence of its trustworthiness.

Table 4: Summary of the classification report for the LSP classifiers.

Classifier	Classes	True Positive Rate (%)	False Negative Rate (%)	Positive Predictive Value (%)	False discovery rate (%)	Accuracy (%)
Ensemble Boosted Trees	Head	73	27	73	27	**75.6**
	Lower arm	67	33	77	23	
	Lower leg	67	33	71	29	
	Torso	80	20	80	20	
	Upper arm	87	13	76	24	
	Upper leg	80	20	75	25	
KNN	Head	100		79	21	74.4
	Lower arm	53	47	50	50	
	Lower leg	60	40	100		

	Torso	73	27	79	21	
	Upper arm	80	20	80	20	
	Upper leg	80	20	71	21	
Linear SVM	Head	87	13	100		78.9
	Lower arm	80	20	86	14	
	Lower leg	60	40	100		
	Torso	100		52	48	
	Upper arm	73	27	92	08	
	Upper leg	73	27	85	15	
Quadratic SVM	Head	62	38	100		78.6
	Lower arm	92	08	92	08	
	Lower leg	75	25	100		
	Torso	100		48	52	
	Upper arm	77	23	83	17	
	Upper leg	69	31	100		

Our assessment was conducted on public benchmark datasets (MPII and LSP), which are commonly utilized for pose estimation work. Consistent with existing research, we have considered joint-level classification accuracy, as this offers a finer-grained and harder assessment of model performance. In fact, numerous state-of-the-art works also provide results in terms of joint-level and not full-pose reconstruction, so our assessment is conformant to accepted practice.

In order to enhance statistical rigor and reproducibility, we have included a k-fold cross-validation (k = 5) protocol and are now using mean accuracies with standard deviations. This gives a better estimation of the performance of the model and avoids our results being skewed towards any specific data split. Although our stated accuracies (74–79%) may seem modest, it should be pointed out that these are competitive in relation to work being done on the same datasets, where joint-level evaluation is used in a majority of cases.

In addition, our pipeline was constructed as a modular structure, and although the present work focuses on joint-level classification, it can be easily generalized towards full-pose accuracy measures (e.g., PCKh) in subsequent work. This puts our contribution into the larger research context of human pose estimation while ensuring reproducibility and conformance to standard evaluation methodologies.

4. Ablation study and Module Contribution

For a better understanding of the contribution of every module, we have performed an ablation study in qualitative terms. The preprocessing stage, i.e., morphological segmentation, was responsible for removing the background noise and maintaining the structure of the human body and resulted in more accurate pose regions. The use of Weber's Law-driven contrast enhancement also made pose landmarks more separable and further enhanced joint boundaries,

improving the feature separability. In feature extraction, VGG16 was the most essential part, as its deep features adequately captured pose-related variations and outperformed conventional descriptors. Lastly, the utilization of ensemble classifiers proved the generalization power of the proposed pipeline: on the MPII dataset, the best accuracy of 74.4% was realized with Linear Discriminant Analysis, whereas on the LSP dataset, Linear SVM achieved 78.9% accuracy. Together, these modules show that the combination of preprocessing, contrast enhancement, deep feature extraction, and optimized classifiers leads to a strong and efficient human pose estimation system.

5. Conclusion

This paper offers a strong human pose estimation architecture that utilizes morphological segmentation to efficiently pre-process, Weber's Law-based contrast improvement, and deep feature extraction using the VGG16 convolution neural network. Features were tested using various conventional machine learning classifiers on two difficult datasets: MPII and LSP. The Linear Discriminant classifier in the MPII dataset reached its best accuracy of 74.4%, demonstrating the power of linear modeling for recognizing pose-related patterns from VGG16 features. Ensemble and discriminant classifiers also worked well, with impressive performances from Ensemble Subspace KNN and Quadratic Discriminant classifiers. The system performed even better on the LSP dataset, where Linear SVM registered 78.9%, followed closely by Quadratic SVM at 78.6%, showing the strength of SVM-based solutions on this dataset. These results validate that mixing traditional image enhancement methods with deep learning–inspired feature extraction and wisely chosen classifiers can provide effective pose classification systems.

Table 5: Benchmark comparison of human pose estimation techniques on standard datasets.

Ref	Methods	Datasets	Accuracy (%)
[22]	SVM (MSR3D)	MSR3D	72.5
[23]	CNN	COCO	69.7
[24]	Graph Convolutional Network (GCN)	MPII / COCO	69.7
[25]	CNN + RNNs	MPII / LSP / FLIC	73.4
[26]	PRHP (Probabilistic Human Pose)	MPII	74.0
Proposed Method (MorphSeg + Weber +VGG16 + SVM)		MPII / LSP	74.4 MPII 78.9 LSP

Additionally, Table 5 shows a benchmark comparison of some current pose estimation methods. End-to-end learning and large-scale training-based deep CNN and GCN-based models on MPII, LSP, and COCO datasets tend to have higher accuracies (74–80% and higher) because of probabilistic refinement strategies and spatial–temporal dependencies modeled. For instance, CNN–RNN hybrids and PRHP-based models perform better than conventional classifiers through modeling of spatial–temporal dependencies and probabilistic refinement strategies. Comparatively, our suggested method, though slightly lower in absolute performance, has competitive performance (74.4% on MPII and 78.9% on LSP). Notably, our approach is computationally light and interpretable, providing a balance between efficiency and accuracy. This confirms the efficiency of combining morphological segmentation, Weber's Law-based enhancement, and VGG16 feature extraction with classical classifiers as an alternative to more sophisticated end-to-end deep architectures that are computationally expensive.

6. References

1. Song, L., et al., *Human pose estimation and its application to action recognition: A survey.* Journal of Visual Communication and Image Representation, 2021. **76**: p. 103055.

2. Kumar, P., S. Chauhan, and L.K. Awasthi, *Human pose estimation using deep learning: review, methodologies, progress and future research directions.* International Journal of Multimedia Information Retrieval, 2022. **11**(4): p. 489-521.
3. Chen, H., et al., *2D Human pose estimation: A survey.* Multimedia systems, 2023. **29**(5): p. 3115-3138.
4. Gamra, M.B. and M.A. Akhloufi, *A review of deep learning techniques for 2D and 3D human pose estimation.* Image and Vision Computing, 2021. **114**: p. 104282.
5. Li, T. and H. Yu, *Visual–inertial fusion-based human pose estimation: A review.* IEEE Transactions on Instrumentation and Measurement, 2023. **72**: p. 1-16.
6. Nguyen, H.-C., et al., *Unified end-to-end YOLOv5-HR-TCM framework for automatic 2D/3D human pose estimation for real-time applications.* Sensors, 2022. **22**(14): p. 5419.
7. Sigal, L., *Human pose estimation*, in *Computer Vision: A Reference Guide*. 2021, Springer. p. 573-592.
8. Samkari, E., et al., *Human pose estimation using deep learning: A systematic literature review.* Machine Learning and Knowledge Extraction, 2023. **5**(4): p. 1612-1659.
9. Bourdev, L. and J. Malik. *Poselets: Body part detectors trained using 3d human pose annotations*. in *2009 IEEE 12th international conference on computer vision*. 2009. IEEE.
10. Dalal, N. and B. Triggs. *Histograms of oriented gradients for human detection*. in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. 2005. Ieee.
11. Felzenszwalb, P.F. and D.P. Huttenlocher, *Pictorial structures for object recognition.* International journal of computer vision, 2005. **61**: p. 55-79.
12. Aidoo, E., et al., *Cofopose: Conditional 2d pose estimation with transformers.* Sensors, 2022. **22**(18): p. 6821.
13. Kulkarni, S., et al., *Poseanalyser: A survey on human pose estimation.* SN Computer Science, 2023. **4**(2): p. 136.
14. Chung, J.-L., L.-Y. Ong, and M.-C. Leow, *Comparative analysis of skeleton-based human pose estimation.* Future Internet, 2022. **14**(12): p. 380.
15. Srividya Inampudi, Romik Amipara, and R. Kokate, *Human Pose Detection and Estimation.* IJCRT, 2020. **8**(9 September 2020).
16. Ding, W., et al., *Human posture recognition based on multiple features and rule learning.* International Journal of Machine Learning and Cybernetics, 2020. **11**: p. 2529-2540.
17. Chen, H., et al., *2D Human Pose Estimation: A Survey.* arXiv, 2022. **1**(15 Apr 2022).
18. Kevin, M.A., et al., *Advanced System for Enhancing Location Identification through Human Pose and Object Detection.* Machines, 2023. **11**(8): p. 843.
19. Li, R., et al., *Human pose estimation based on efficient and lightweight high-resolution network (EL-HRNet).* Sensors, 2024. **24**(2).
20. Li, Z., et al., *Lightweight 2D Human Pose Estimation Based on Joint Channel Coordinate Attention Mechanism.* Electronics, 2023. **13**(1): p. 143.
21. Ghasemi-Naraghi, Z., A. Nickabadi, and R. Safabakhsh, *Pose Scoring Model for Refining Multi-Person Poses.* 2025.
22. Wang, J., et al. *Mining actionlet ensemble for action recognition with depth cameras*. in *2012 IEEE conference on computer vision and pattern recognition*. 2012. IEEE.
23. Lin, T.-Y., et al. *Microsoft coco: Common objects in context*. in *European conference on computer vision*. 2014. Springer.
24. Yan, S., Y. Xiong, and D. Lin. *Spatial temporal graph convolutional networks for skeleton-based action recognition*. in *Proceedings of the AAAI conference on artificial intelligence*. 2018.
25. Song, S., et al. *An end-to-end spatio-temporal attention model for human action recognition from skeleton data*. in *Proceedings of the AAAI conference on artificial intelligence*. 2017.

26. Chen, Y., et al. *Adversarial posenet: A structure-aware convolutional network for human pose estimation*. in *Proceedings of the IEEE international conference on computer vision*. 2017.