Yoga Pose Detection Using Distance Transform and Convolutional Neural Networks

1st Bora Mert Şahin
TOBB University of Economics and Technology
Biomedical Engineering
Ankara , Turkey
boramertsahin@etu.edu.tr
https://orcid.org/0000-0003-3461-8573

***Abstract*—Creating basic skeleton structure from images of a person can be done using machine learning methods easily. But it is a difficult task to do using only image processing methods. There are many factors that effect the skeletonization process such as lighting and background environment.**
This study suggests the use of Distance Transform to create a simple, 1 pixel wide skeleton of a person. Skeleton images were obtained from images of people performing 5 different yoga positions. A Convolutional Neural Network (CNN) was used to classify yoga positions from original images and skeleton images seperately. CNN model trained with skeleton images showed higher test and validation accuracy compared to the model trained with original images.

***Index Terms*—Skeletonization, Computer Vision, Body detection, Machine Learning**

I. Introduction

It is difficult to detect human bodies using a 2D image since human body has a complex structure and people wear clothes with different fit and texture. One method for making the task of human body detection is called skeletonization. This method creates a stick-figure-like structure from the image of a human, eliminating the variations caused by physical appearance of different people.
There are different methods for human image skeletonization. A basic, star shaped skeleton can be used to represent human images [1] [2]. Some methods include the use of 3D image data obtained from devices such as Kinect [3]. There is also a study for using machine learning to create a 3D model of a human body and match that model to the 2D image of a person [4].
This study uses the gradient of the distance transform [5] [6]. Distance transform creates a grayscale image corresponding to the distance of the each pixel of the foreground object to the background. When applied to an image of a person with black background, distance transform creates a thin version of the person in the image as seen in Fig 1

This study aims to create a skeletonized image of a person doing yoga poses and use the result image to investigate the effect of skeletonization in image classification using CNN compared to original images.

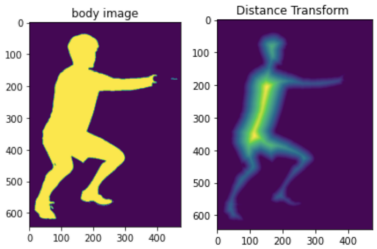

Fig. 1. Binary image with black background and white human silhouette (left), distance transformed image (right).

II. Materials and Methods

A. Dataset

"Yoga Poses Dataset" from Kaggle was used for this study [7]. Dataset consists of 5 different yoga poses with total number of 1551 images (1081 train, 470 test images). Some of the images were removed from the dataset due to incompatibility and repeating images. After removal, dataset consisted of 734 train and 186 test images.

B. Image Processing

Background of the images were removed using the Python tool "rembg" which uses U2-Net, an architecture designed for salient object detection (SOD) [8]. Result image is converted to a binary image by setting the background color as black and every pixel belonging to foreground object as white.
Euclidean Distance Transform (EDT) is applied to the binary image. EDT simply measures the distance of white pixels to the closest black (edge) pixel using the equation below:

$$D_{Euclid} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

Fig. 2. Simple example of Euclidean Distance Transform [9].

Local Maximum Points (LMP): In the distance transform image, each pixel and neighboring pixels are inspected. If the value of center pixel is larger than all neightboring pixels, that pixel is marked as LMP.

Critical Points (CP): Calculating the gradient of the distance transform (ΔDT) results in even thinner skeleton image with minimum values in vertical and horizontal parts of the skeleton. Finding the local minimum points with similar method to the LMP, and taking the LMP points where ΔDT is the local minimum gives the Critical Points of the skeleton as seen in Fig 3

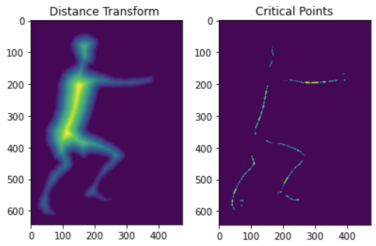

Fig. 3. Distance transform (left), critical points of skeleton (right).

C. Classification

A CNN model was used for classification in this study. CNNs are primarily used to solve difficult image-driven pattern recognition tasks. This study hypothesizes that using skeletonized images would create more simple patterns for detection of human poses. We investigated this hypothesis by training a CNN model with original and skeleton images seperately and compare the performance of the model.
Structure of the model can be seen in Fig 4. 20% of the training images were split for validation.

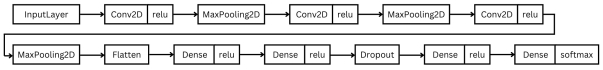

Fig. 4. CNN model.

III. Results & Discussion

Train and test accuracy results of both original images and skeleton images can be seen in Table I

TABLE I
TABLE 1. TRAIN AND TEST ACCURACY RESULTS OF ORIGINAL AND SKELETON IMAGES.

Accuracy	Image Type	
	Original	***Skeleton***
Train (%)	87.96	96.07
Validation (%)	73.45	93.06
Test (%)	86.80	97.24

As seen in Figure 5, model trained with skeleton images had minor errors with the test data while the model trained

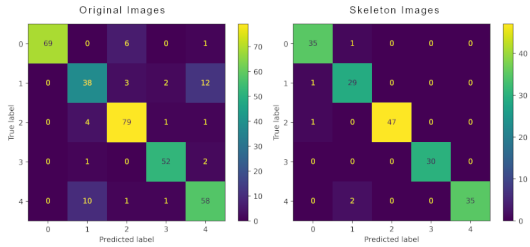

Fig. 5. Confusion matrix of test data evaluation.

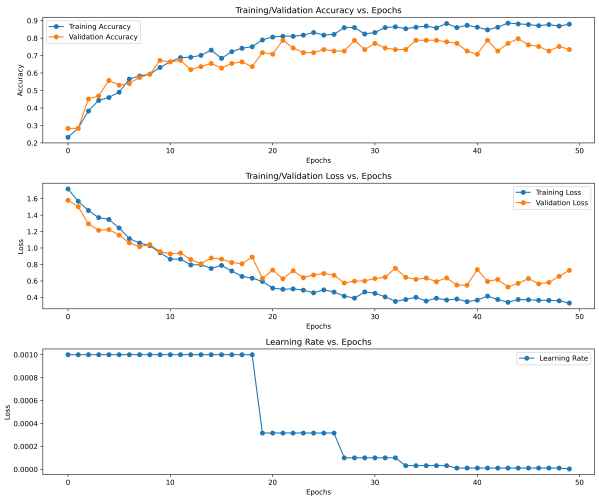

Fig. 6. Comparison of Training/Validation accuracy, loss and learning rate vs. epochs for Original Images.

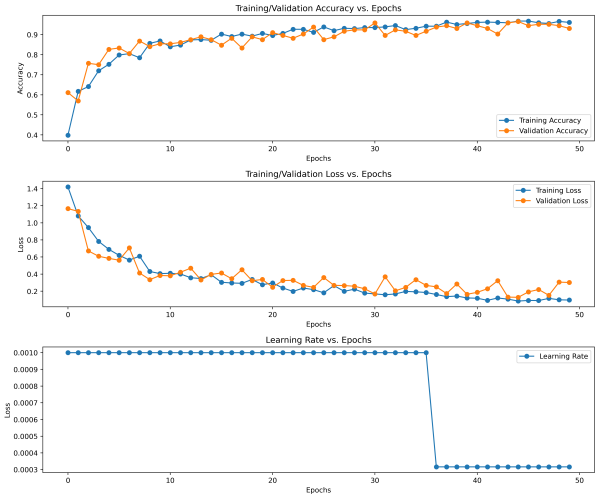

Fig. 7. Comparison of Training/Validation accuracy, loss and learning rate vs. epochs for Skeleton Images.

with original images had higher number of error in general.

CNN model trained with original images had higher training accuracy than validation accuracy (Fig 6), while the model trained with skeleton images had these two values very close to each other (Fig 7. Higher training accuracy compared to the validation accuracy can indicate the risk of overfitting. Since the skeleton model had closer values, it can be less prone to this problem.

CONCLUSION

This study aimed to create a new dataset using skeletonization with distance transform for CNN models to get more accurate results. As seen in results, same CNN model trained with skeletonized images showed higher classification performance compared to the original images of people doing yoga poses.

In order to consider this method a valid method, this study should be repeated with a larger dataset.

Currently, the skeleton image obtained from the critical points is just some points scattered around the place where the skeleton should be. In future works, these points can be combined to create a complete stick figure. Angle and position data of the stick figure can be turned into a feature matrix and machine learning models other than CNN can be used to classify human images.

REFERENCES

[1] Chen, H. S., Chen, H. T., Chen, Y. W., Lee, S. Y. (2006, October). Human action recognition using star skeleton. *In Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks (pp. 171-178).*

[2] Fujiyoshi, H., Lipton, A. J., Kanade, T. (2004). Real-time human motion analysis by image skeletonization. *IEICE TRANSACTIONS on Information and Systems, 87(1), 113-120.*

[3] Schwarz, L. A., Mkhitaryan, A., Mateus, D., Navab, N. (2012). Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing, 30(3), 217-226.*

[4] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10975-10985).*

[5] Ding, J., Wang, Y., Yu, L. (2010, March). Extraction of human body skeleton based on silhouette images. *In 2010 Second International Workshop on Education Technology and Computer Science (Vol. 1, pp. 71-74). IEEE.*

[6] Niblack, C. W., Gibbons, P. B., Capson, D. W. (1992). Generating skeletons and centerlines from the distance transform. *CVGIP: Graphical Models and image processing, 54(5), 420-437.*

[7] Yoga Poses Dataset. (2023). Retrieved 13 April 2023, from *https://www.kaggle.com/datasets/niharika41298/yoga-poses-dataset*

[8] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition, 106, 107404. doi: 10.1016/j.patcog.2020.107404*

[9] Distance Transform of a Binary Image - MATLAB amp; Simulink. Available at: https://www.mathworks.com/help/images/distance-transform-of-a-binary-image.html (Accessed: April 13, 2023).