https://doi.org/10.2352/ISSN.2470-1173.2018.09.IRIACV-204


# A 3D Guitar Fingering Assessing System Based on CNN-Hand Pose Estimation and SVR-Assessment

***Zhao WANG and Jun OHYA***
***Department of Modern Mechanical Engineering, Waseda University, Tokyo, Japan***

## Abstract

*This paper proposes a guitar fingering assessing system based on CNN (Convolutional Neural Network) hand pose estimation and SVR (Support Vector Regression) evaluation. To spur our progress, first, a CNN architecture is proposed to estimate temporal 3D position of 16 joints of hand; then, based on a DCT (Discrete Cosine Transform) feature and SVR, fingering of guitarist is scored to interpret how well guitarist played. We also release a new dataset for professional guitar playing analysis with significant advantage in total number of video, professional judgement by expert of guitarist, accurate annotation for hand pose and score of guitar performance. Experiments using videos containing multiple persons' guitar plays under different conditions demonstrate that the proposed method outperforms the current state-of-art with (1) low mean error (Euclid distance of 6,1 mm) and high computation efficiency for hand pose estimation; (2) high rank correlation (0.68) for assessing the fingering (C major scale and symmetrical excise) of guitarists.*

## Introduction

Recently, with the development of computer vision, the automatic guitar fingering teaching system has been attracting lot attentions of academic research [4−9]. The reason of guitar fingering need to be assessed by computer vision rather than audio is that, a large number of possible fingering alternatives exist just for a same single set of notes. It means even though audio signal may sound right, the fingering, which is much more important than sounding maybe wrong [1,3]. However, there are some challenging points in computer vision to solve these problems: first, as the hand of human is flexible, fast-moving in motion when playing guitar, hand region is hard to be segmented concerning complex background, different illumination situation and so on; also compared with other pose tracking or pose estimation problems such as pedestrian tracking or human pose estimation, because of the flexibility and self-occlusion in guitar playing context, hand pose of guitarist is difficult to be estimated or tracked; furthermore, action assessing, which could automatically evaluate or score the quality of action is quite different than action recognition, and only few promising result of action-assessing algorithm is explored and still a long way to rivaling the performance of expert judges.

The guitar fingering assessing system could be interpreted into three steps: (1) hand region segmentation, (2) hand pose estimation and (3) fingering assessment. Related works of guitarist supporting systems solve the problems shown as below [2]: Motokawa and Saito [5] built a system called Online Guitar Tracking that supports a guitarist by using augmented reality. This is done by showing visual aid information (i.e. the virtual fingers model) on a real stringed guitar and this becomes an aid to learning to play the guitar. Online Guitar Tracking uses augmented reality to detect the guitar so that the player can learn how to hold the strings of the guitar by overlapping the player's hand onto a manual model. Scarr and Green [6] proposed an algorithm that uses a markerless approach to successfully locate a guitar fretboard in a webcam image by using Hough Transform, and detect the skin color by only using a single RGB color threshold. Burns and Wanderley [7] detected the positions of fingertips for the retrieval of the guitarist fingering without markers. They used the circular Hough transform to detect fingertips. By fixing a camera on the guitar neck, the guitar neck and the camera are relatively static, but the fixed camera brings a lot of inconveniences to guitar players. Kerdvibulvech and Saito [8-10] proposed a series of novel approaches to detect the position of the player's fingers such as stereo cameras are used to compute the 3D positions of fingers using the color markers attached on fingertips, and template matching is used to localize the fingertip positions etc. All the research fix only partial function in guitar fingering teaching such as chord recognition, guitar neck tracking and finger detection. Thus they show more faultiness than novelties: (1) some of them use inconvenience tool, such us color markers, neck-fixed camera and ARTag [5,7,8], and it brings a lot of inconveniences to the guitarist; (2) Instead of tracking or estimating finger pose to obtain the comprehensive fingering of guitarist, some of them only detect finger [6,7,9] in certain frames; (3) all of them are extremely sensitive to background and illumination, therefore hard to segment hand area from back ground using only threshold of RGB or RGB-D value whether trained or not, also difficult to estimate hand pose because it is built on the work of hand region segmentation; (4) none of the related works achieves the final goal of the topic: fingering assessment. As we mentioned before, each work is built on other's result, for instance, hand pose estimation is built on the result of hand region segmentation, once the result of former step is not accurate and robust enough, it is extremely hard to further research, and we believe this is the main reason that related works cannot achieve the fingering assessment. This is also the main reason why we use deep learning-based methods in this paper while concerning the versatility and the accuracy of deep learning structure.

In this paper, (1) we propose a data-driven CNN framework that output the 3D position of joints of guitarist frame by frame; (2) we adopt SVR to automatically assess how well guitarist perform in video by extracting the spatio-temporal hand pose features of guitarist and estimating a regression model that predicts the scores of guitar playing.

We also introduce a new dataset for guitar playing analysis as benchmarks of hand pose such as NYU [19] ICML [20] are restricted in terms of number of annotated images, annotation accuracy, articulation coverage, and variation in hand shape (bone length) and viewpoint [18]. Our dataset makes significant advantage in total number of video, professional judgement by expert of guitarist, accurate annotation for hand segmentation, hand pose and score of guitar performance.

In the remainder of this paper, we first introduce the CNN based hand pose estimation in Section 2. Then, we present the fingering assessing in Section 3. Furthermore, we introduce our dataset and evaluate our work respectively in Section 4. Last, we conclude the paper in Section 5

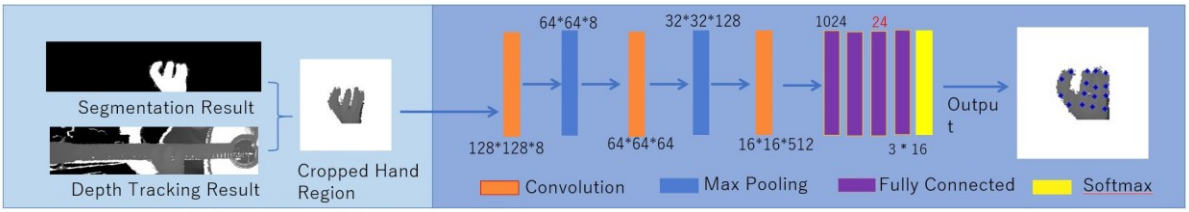


***Figure 1.*** *Process Outline of FCN-based[17,19] Hand Segmentation:* ***a*** *is the 3D Input Video of Guitar Playing;* ***b*** *is the input of the CNN for hand segmentation;* ***c*** *is FCN Structures;* ***d*** *is the Pixel-wised Output of Hand Segmentation.*

## Hand Pose Estimation

### *Hand Region Extraction*

The input of our hand pose estimation is the FCN based segmentation result (under reviewing of another journal paper). After we segment the hand region, we crop the hand region to a 128*128 pixel region like recent deep learning -based approach [13,14] do. However, unlike they assuming the hand is the close object to the object, we use our FCN-based segmentation to crop the hand region automatically: we extract from the depth input and the segmentation mask shown in Fig.1, the centroid of the cropped hand region image is also the centroid of the hand. We implement it by using the segmentation result of FCN-based method (the binarized image), and we trace the longest contour on the segmentation result to get the hand region area. Furthermore, we normalize depth values to [-1,1]: the deeper scene in the cropped image, i.e background is set to 1 (white area in cropped image of Fig.1), and the nearest pixel in hand region is set to -1.The reason we do this normalization is that we need to assure our work is invariant to (1) the distance between camera and hand, (2) the position of the hand in the segmentation result. The process is shown in light blue area of Fig.1.

### *CNN based Hand Pose Estimation*

Related works [13,14] always suggest the deeper, or a multi-scale approach works better. However, in this section, we propose a simple network, which only contains 3 convs and 3 fully-connects works almost equally accurate as the state-of-arts, while outperforms them in better training time efficiency.

Our proposed network is shown in dark region of Fig.2. After we extract the hand region to a 128*128 region, first three Convs layers and two max-pooling layers output 512 channels of feature maps; then we use two fully-connected layers with 1024 notes respectively; furthermore, instead of directly estimating the 3D position of each joint, we predict a lower parameter space as there is a strong relation between each joint of a hand concerning the physical constraint of hand. Also, related work [13] has shown a low dimensional embedding of hand parameter is sufficient to parameterize the hand's 3D pose. Thus, in our case, we implement a fully-connected layer with only 24 notes (red number is Fig.2) after the two 1024-note-FC layers; finally, a fully-connected layer with 3*J (J is the number of the joints, in our case, J=16) notes output the 3D position of hand pose.

## Fingering Assessment

In this section, we present the fingering assessment for evaluating the quality of guitarist's fingering from image sequence. Generally, we formulating a regression model of SVR using the feature of spatio-temporal DCT (discrete cosine transformation) to score how well guitarists play in a video. We built our work on related work [15], and extend it to 3D spaces.

### *Features of fingering*

Given a video of guitar playing, after we estimating the hand pose of guitarist frame by frame, we formulate the hand pose as:

$$q^{(j)}(t) = p^{(j)}(t) - p^{(0)}(t) \quad j \in (1,2 \dots J) \qquad (1)$$

where $J$ is the total number of the finger joint, $t$ is the frame index of the video, $p^{(0)}(t)$ is the index finger's tip at frame $t$. The result we do this normalization is we let our feature be invariant to any input. Then we transform the normalized position of joints from spatial domain to frequency domain:

$$Q^{(j)} = A\, q^{(j)} \qquad (2)$$

$$A_{mn} = \frac{1}{\sqrt{M}} \quad when\ m = 0, 0 \le n \le M-1$$

$$A_{mn} = \sqrt{\frac{2}{M}} cos\frac{\pi(2n+1)m}{2M} \quad when \quad 1 \le m \le M-1,\ 0 \le n \le M-1 \qquad (3)$$

where $A_{mn}$ is $m$th row $n$th column in $A$ of Eq.(2); $M$ is number of total rows of cosine matrix. We compute the features for every joint for 3D (x, y, z) component respectively and concatenate them all to final feature vector $\alpha = |Q|$ , which is absolute value of the result of cosine transform.

### *Scoring Fingering based on SVR Model*

We implement our guitar fingering assessing system as a supervised regression SVR model. Note feature vector $\alpha = |Q|$ is the frequency feature of hand pose, and each video of guitar playing is a single feature vector which horizontal component is the joint index of the guitarist, while vertical one is the first k component of cosine transform. The ground truth of the scoring of each video is obtained by specialist of guitar playing. We implement a linear support vector regression (L-SVR) to predict the score of guitar playing by using lib-svm [15,16]. Any details of SVR could be found at [17].

## Evaluation

In this section, first we introduce our dataset used for guitar fingering analysis and we evaluate our work by evaluating two parts of our work: CNN-based hand pose estimation and SVR-based fingering assessment respectively. Because each part of our works highly depends on the results of others, i.e. the inputs of SVR-based guitar fingering assessing is the output of CNN-based hand pose estimation, we believe this two-step based evaluation could assess our work comprehensively.

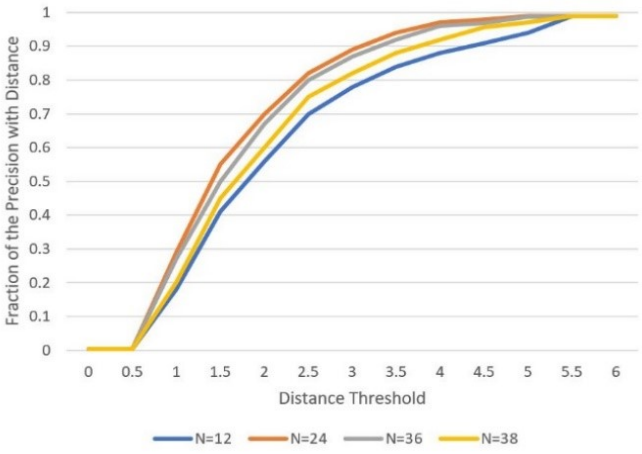


***Figure 2**. Self-Comparison. The horizontal axis indicates the thresholds of mean error of tracking, while the vertical one means the precision of tracking when each threshold on the horizontal axis is set. N=24 achieves best performance within our work.*

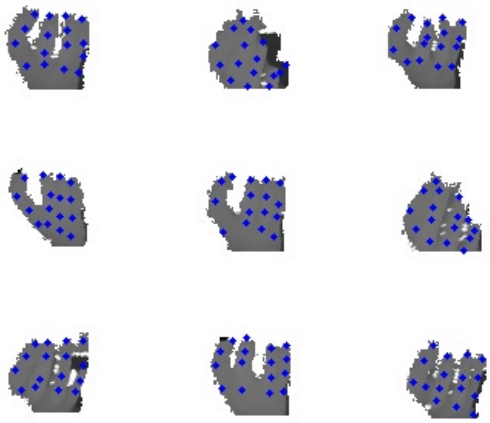

***Figure 3**. Examples of Hand Pose Estimation on an Image Sequence: the blue circles indicate the positions of 16 joints estimated by our method (4 joints for each finger,and hand pose of 4 fingers (index finger, middle finger, ring finger and little finger) are estimated).*

## *Data-set*

Existing benchmarks of hand pose such as NYU [19] and ICML [20] are restricted in terms of number of annotated images, annotation accuracy, articulation coverage, and variation in hand shape (bone length) and viewpoint [18]. Therefore, training on benchmarks is not effective to solve our own problems as they are all overfitted to own dataset. In this section, we introduce our dataset that makes significant improvement in guitar playing analyze problem with the advantages of: (1) all the image sequences are real guitar playing scene including 70 videos with nearly 8000 images for color images and depth images respectively (taken by Microsoft Kineck V2); (2) all the data are taken ranging from expert guitarist to beginner to fulfill the purpose of our guitar fingering analyzing; (3) all the data are labelled by expert guitar players including scores of performance, pixel-wise segmentation annotation and accurate hand pose position for 16 joints of guitarist.; (4) all the data are taken under highly complex background and different illumination situation to let us make a fair evaluation.

## *CNN-based Hand Pose Estimation*

We apply CNN-based Hand Pose Estimation on 524 images in our own dataset. As we described before, we first crop the hand region area centered at the centroid of hand, resize it to 128*128 pixels hand normalize the depth value to [-1,1]. We implement with the Caffe module on a NVIDIA Titan Black GPU with CuDNN V4 acceleration. We use stochastic gradient descent (SGD) with a mini-batch size of 12. The learning rate is fixed to 0.0001, and we train the variants until the training loss converges. In the meanwhile, we use a weight decay of 0.0005 and a momentum of 0.9. we manually label the position of each 16 joints of cropped hand images to assure the accuracy of the annotation of the training data.

The performance is evaluated by two metrics: (a) per-joint mean error of Euclidean distance (in millimeters) and (b) percentage of frames in which all errors of joints are below a threshold. States-of-arts [13,14,21] all calculate the (b) as this metric is generally regarded very challenging, as a single dislocated joint deteriorates the whole hand pose [13,21]. First, we self-compare our net with baseline and different ensemble settings on our net structure. The self-comparison result based on Metric (a) and (b) are shown in Table 1 and Fig.2 respectively**.** Then we compare our work with several state-of-the-arts [13,21] methods on our dataset not only in accuracy of estimation (Table 1) but also in time efficiency of training and testing (Table 2). Our work shows a competitive accuracy with state-of-arts [13,21], but outperform them in time efficiency for both training and testing. An example of hand pose estimation result on a video is shown in Fig.3.

## *SVR-based Fingering Assessment*

In our fingering assessment evaluation, we choose 2 kinds of music pieces, which are the most frequently daily practices for guitarist: (1) ***C*** major scale on first fret and (2) symmetrical excise as the object of fingering assessment. Both of the two excises are fundamental, classic practices but best way to improve dexterity, speed, strength and stamina to help you overcome obstacles and become a better guitar player [22]. Each music piece contains 50 videos by Microsoft Kinect, and totally there is almost 4000 frames

**Table 1. Per-joint mean error of hand pose estimation for Self-Comparison (First Four Rows) and Comparison with State-of-arts (Last Two Rows).** I, M, R, L Indicate Index Finger, Middle Finger, Ring Finger and Little Finger; 1, 2, 3, 4 Indicate Finger Joint from Tips to Root.

| | I.1 | I.2 | I.3 | I.4 | M.1 | M.2 | M.3 | M.4 | R.1 | R.2 | R.3 | R.4 | L.1 | L.2 | L.3 | L.4 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***N*** = 12 | 12.1 | 4.5 | 5.7 | 3.4 | 11.4 | 10.5 | 6.9 | 7.4 | 10.6 | 7.7 | 6.5 | 7.5 | 11.0 | 12 | 4.8 | 4.4 | 7.9 |
| ***N*** = 24 | 13.2 | 4.0 | 4.1 | 2.9 | 12.1 | 9.6 | 5.1 | 6.6 | 11.4 | 4.8 | 4.2 | 3.5 | 9.2 | 3.8 | 4.4 | 4.2 | **6.1** |
| ***N*** = 36 | 13.3 | 4.3 | 4.4 | 4.0 | 13.5 | 7.6 | 5.0 | 6.7 | 12.8 | 5.8 | 5.2 | 3.7 | 10.3 | 3.9 | 2.8 | 4.7 | 6.5 |
| ***N*** = 48 | 14.8 | 5.0 | 5.0 | 4.7 | 13.9 | 9.6 | 4.8 | 4.7 | 11.0 | 7.7 | 5.2 | 3.9 | 12.0 | 4.4 | 4.7 | 5.5 | 7.3 |
| REN[21] | 12.9 | 4.0 | 4.4 | 3.1 | 9.5 | 8.3 | 6.6 | 6.7 | 12.1 | 3.0 | 3.9 | 3.5 | 8.2 | 4.4 | 3.7 | 3.3 | **6.1** |
| D.P[13] | 13.3 | 4.1 | 4.2 | 3.7 | 9.6 | 7.0 | 6.3 | 6.4 | 10.6 | 4.1 | 4.9 | 3.2 | 10.5 | 3.2 | 4.5 | 3.6 | 6.2 |

**Table 2. Comparison of Time Efficiency.**

| | Training Time (h) | Testing time per Image (ms) with GPU |
|---|---|---|
| Our Work | **4** | **0.19** |
| REN [21] | 21 | 0.84 |
| Deep Prior [13] | 16 | 0.56 |

of color and depth respectively for each music piece. Each video is scored ranging from 0 to 100 (full mark) by experts of guitar playing. We cross validate our work by randomly selecting 40 videos as training, 10 as testing for 10 times.

We self-compare our work against several baselines: (1) we utilize three features: a. 3D-DCT as our proposed, b. 3D Discrete Fourier Transform (DFT), c.3D Space-time interest points (STIP); (2) we compare the ridge regression with our proposed SVR with three features mentioned in (1) respectively. Table 3a and 3b show the detailed comparison results on C major scale and symmetrical excise, respectively. We use mean rank correlation to evaluate our work like states-of-arts did [15], and mean rank correlation is defined as follows:

$$\rho = 1 - \frac{6\sum_1^N (x_i - y_i)^2}{n(n^2-1)} \qquad (4)$$

where $N$ is total number of the testing data, $x_i$ is the regression score of $i$-th data, $y_i$ is the ground truth score of $i$-th data. Our proposed method (3D-DCT+SVR) outperform others with a mean rank correlation of 0.68 for C major scale and symmetrical excise (0.67 in Table 3a and 0.69 in Table 3b.).

In the meantime, we compare our proposed fingering assessment with the AQA [15], which is the only work for automatically evaluating human action on 2D video to our knowledge. From Table 3a and 3b, we figure out our proposed method outperform it with the mean rank correlation 0.68 (the mean rank correlation of AQA [15] is 0.415 tested on our dataset of ***C*** major scale and symmetrical excise). Two examples of assessing the guitar fingering is shown in Fig.4: in Fig.4a, the predicted score of the guitar playing is 88.7 while the ground true labeled by expert guitarist is 90; in Fig.4b, the predicted score of our system is 21.7, while the ground truth is 20.

## Conclusion

In this paper, we have proposed a guitar fingering assessing system based on CNN (Convolutional Neural Network) hand pose estimation and SVR (Support Vector Regression) evaluation. To spur our progress, first, an end-to-end, pixel-wise FCN is trained to segment hand area of guitarist; then a CNN architecture is proposed to estimate temporal 3D position of 16 joints of hand; finally, based on a DCT (Discrete Cosine Transform) feature and SVR, fingering of guitarist is scored to interpret how well guitarist played. We also release a new dataset for professional guitar playing analysis with significant advantage in total number of video, professional judgement by expert of guitarist, accurate annotation for hand segmentation, hand pose and score of guitar performance.

*a. An Example of High Score Performance of Symmetrical Excise (Ground true: 90; Our Assessing Result: 88.7)*

*b. An Example of Low Score Performance of C Major Scale (Ground true: 20; Our Assessing Result: 21.7)*

***Figure 4**. Two Examples of Assessing Result on Our Dataset*

**Table 3a. Mean Rank Correlation of Fingering Assessing for C Major Scale** (the first row is the mean ran correlation value of our proposed SVR; while the second row is the mean ran correlation value of Ridge Regression. The different columns indicate the value of mean ran correlation for each different features: DCT, DFT etc.)

| | 3D-STIP | 3D-DCT | 3D-DFT | AQA[15] |
|---|---|---|---|---|
| SVR | 0.33 | **0.67** | 0.45 | 0.34 |
| Ridge Regression | 0.29 | 0.45 | 0.38 | 0.27 |

**Table 3b. Mean Rank Correlation of Fingering Assessing for Symmetrical Excise**

| | 3D-STIP | 3D-DCT | 3D-DFT | AQA[15] |
|---|---|---|---|---|
| SVR | 0.31 | **0.69** | 0.57 | 0.49 |
| Ridge Regression | 0.35 | 0.48 | 0.49 | 0.44 |

Experiments using videos containing multiple persons' guitar plays under different conditions demonstrate that the proposed method outperforms the current state-of-art with (1) low mean error (Euclid distance of 6,1 mm) and high computation efficiency for hand pose estimation; (2) high rank correlation (0.68) for assessing the fingering (C major scale and symmetrical excise) of guitarists.

Remaining issues include (1) the feedback of the system that gives practicing advice to guitar players to help them with their fingering; (2) the explorations of more efficient fingering assessment such as reinforcement learning etc.

## References


[1] Radisavljevic, Aleksander, and Peter Driessen. "Path difference learning for guitar fingering problem." Proceedings of the International Computer Music Conference. Vol. 28. Sn (2004).

[2] Wang, Zhao, and Jun Ohya. "Fingertips Tracking Algorithm for Guitarist Based on Temporal Grouping and Pattern Analysis." Asian Conference on Computer Vision (ACCV). Springer, Cham. pp. 212-226 (2016).

[3] Sayegh,S.I. "Fingering for String Instruments with the Optimum Path Paradigm" Computer Music Journal, vol.13, No. 3, Fall 1989, pp. 76-83 (1989).

[4] Radicioni, Daniele, and Vincenzo Lombardo. "Guitar fingering for music performance." strings. Vol. 40. No. 45 (2005)

[5] Y. Motokawa and H. Saito, "Support system for guitar playing using augmented reality display," in Proceedings of the 2006 Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)- Volume 00. IEEE Computer Society, pp. 243–244 (2006).

[6] Joseph Scarr, Richard Green, "Retrieval of Guitarist Fingering Information using Computer Vision," Image and Vision Computing New Zealand (IVCNZ), 2010 25th International Conference, ISSN : 2151-2191 , pp. 1-7, (2010).

[7] A. Burns, “Visual Methods for the Retrieval of Guitarist Fingering”, Proceeding of the 2006 conference on New interfaces for musical expression ISBN:2-84426-314-3, pp.196-199, (2006).

[8] Chutisant Kerdvibulvech and Hideo Saito, “Real-Time Guitar Chord Estimation By Stereo Cameras For Supporting Guitarists”. In Proceeding of 10th International Workshop on Advanced Image Technology 2007 (IWAIT), pp.147-152, (2007).

[9] Chutisant Kerdvibulvech and Hideo Saito, “Guitarist Fingertip Tracking by Integrating a Bayesian Classifier into Particle Filters”. International Journal of Advances in Human-Computer Interaction (AHCI), pp121-131, (2008).

[10] Kerdvibulvech, Chutisant, and Hideo Saito. "Markerless guitarist fingertip detection using a bayesian classifier and a template matching for supporting guitarists." Proceedings of the 10th ACM/IEEE Virtual Reality International Conference, VRIC ‘08, (2008).

[11] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." arXiv preprint arXiv:1511.00561 (2015)

[12] Long, J., Shelhamer, E., & Darrell, T. “Fully convolutional networks for semantic segmentation”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) pp. 3431-3440, (2015).

[13] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, pages 3316–3324, (2015).

[14] Zhou, Xingyi, et al. "Model-based deep hand pose estimation." arXiv preprint arXiv:1606.06854 (2016).

[15] Pirsiavash, Hamed, Carl Vondrick, and Antonio Torralba. "Assessing the quality of actions." European Conference on Computer Vision. Springer, Cham, (2014).

[16] Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) (2011)

[17] Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. NIPS (1997)

[18] Yuan, Shanxin, et al. "BigHand2. 2M Benchmark: Hand Pose Dataset and State of the Art Analysis." arXiv preprint arXiv:1704.02612 (2017).

[19] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. In TOG, (2014)

[20] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3D articulated hand posture. In CVPR, (2014)

[21] Guo, Hengkai, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. "Region Ensemble Network: Improving Convolutional Network for Hand Pose Estimation." arXiv preprint arXiv:1702.02447 (2017).

[22] http://www.guitarworld.com/three-steps-shred-fundamental-daily-practice-techniques-about-15-minutes.

[23] Zhao, Wang, Ohya, Jun. “Detecting and Tracking the Guitar Neck Towards the Actualization of a Guitar Teaching-aid System”, the... international conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics, No. pp. 187-188, (2015).

[24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, (2014).

## Author Biography.

*Zhao WANG is now a current PhD candidate in the department of MME (Modern Mechanical Engineering) in Waseda University. He got the Bachelor Degree in Sun Yet-sun University in China (2010), and Master Degree in Waseda University (2015). Now he is mainly working on tracking algorithm of Computer Vision and Machine Learning.*

*Dr. Jun Ohya is a professor at the Department of Modern Mechanical Engineering, Waseda University, Japan. He earned his B.S., M.S., and Ph.D. degrees in Precision Machinery Engineering from the University of Tokyo in 1977, 1979, and 1988, respectively. Dr. Ohya is a member of IEEE, IEICE, the Information Processing Society of Japan, etc. His research fields include image processing, computer vision, virtual reality, multimedia, pattern recognition.*