Graduate School of Creative Science and Engineering
Waseda University

博 士 論 文 概 要
Doctoral Thesis Synopsis

論 文 題 目
Thesis Theme

An Automatic Guitar Fingering Assessing System Based on Convolutional Neural Network and Spatio-temporal Support Vector Regression

申 請 者
(Applicant Name)

Zhao	WANG
王	釗

Department of Modern Mechanical Engineering,
Research on Image Engineering

May, 2018

Learning how to play guitar is a very interesting but extremely complicated process, as it requires guitarist to do many techniques at the same time: reading scores, pressing fretboard, sweeping or plucking strings and so on so forth. One of the most difficult skills in guitar learning is called “Fingering”, which is a cognitive process that maps each note on a music score to a fingered position of left hand on guitar fretboard. Consequently, for a score containing n notes, there can exist a maximum of 16^n combinations of <string, fret, finger>. Among them, guitarist needs to execute the notes by selecting few correct ways (in most cases, there is only one correct fingering) to perform beautiful and elegant music. Therefore, the case “sound maybe right but fingering is wrong” exists, and because of this situation happens, it is very likely that beginners of guitar may ignore checking whether his or her fingering is wrong as his or her playing performance sounds not bad. This situation brings a lot of harm to beginners of guitar because it must solicit them to develop a bad fingering habit of guitar playing.

Recently, with the development of computer vision, achieving automatic guitar fingering teaching systems has been attracting a lot of attentions of academic research. However, computer vision based guitar fingering recognition and fingering assessing related works show many drawbacks: (1) some researches require guitarist using supportive tool, such as head-worn camera, color marker, AR Tag, and these tools bring a lot of inconveniences to guitar playing; (3) based on the author’s acknowledgement, there is no previous work on guitar fingering assessing system: no work can assess the fingering of guitar, and give feedback such as general evaluation of the performance to guitarist. (3) some works cannot work under complex circumstance, such as different illumination situations, as they use fixed RGB threshold to segment hand region of players concerning the diversity of human skin appearance, complex background and etc.

On the other hand, the current and the most advanced computer vision-based algorithms related to this thesis such as object tracking algorithm (for guitar neck tracking), hand region segmentation (for guitarist hand region segmentation), multiple targets tracking algorithm and hand pose estimation (for track or estimate fingering of guitarist), human action assessing algorithm (for fingering assessing) also show the weakness when they are implemented in the system, for instance, most of the state-of-arts cannot handle the problems of (1) guitar neck is occluded by the hand of guitarist during playing, (2) Self-occlusion of the hand of the guitarist, (3) training based assessing without human intervention, which are frequently happened and very important factors of the purpose of the thesis.

Towards the actualization of the system, in this thesis, three modules are proposed in this thesis as follows:

(1) Guitar Neck Tracking module:

For the guitar neck tracking, after inputting the video of guitar playing, SIFT feature points are detected on every frame as SIFT Feature is invariant to rotation, illumination and scale changes in images; then a KD-tree searching based algorithm is utilized to match the SIFT features between the first frame and any other frame of input videos; furthermore, a modified version of RANSAC (Random Sample Consensus) to overcome the occluded SIFT issue: SIFT feature is overlapped and occluded by fingers of guitar players during guitar playing. The proposed modified RANSAC filters out and eliminates the mis-matched feature points due to the occluded

SIFT issue, and then recovery all the mis-matched feature; finally, to suppress the effect of the guitar neck motion, the tracked guitar neck area on every frame is projected to a new image sequence that guitar neck area is always projected to the center of the new image sequence. Owing to this projection, no matter how the guitar player shakes or swings the guitar neck while playing, the neck area on every frame is always projected to the center of the new image sequence to facilitate analyzing the fingering.

Experiments using 50 videos of guitar playing with nearly 300 frames of the color images (also 300 frames of depth) of different guitar plays under different conditions show promising results of the proposed method: the total mean tracking error is only 4.2 mm and variance is 1.5 mm for the four tracked corners of the guitar fretboard is obtained. Besides, experiments also show the robustness over the rotation and translation movement: the limitation angle of the rotation using the right hand of guitarist as the pivot is 20 degrees; the limitation angle of the upward rotation using the center line of guitar neck as the pivot is 8 degrees, the limitation angle of the rotation using the horizontal center of guitar neck as the pivot is 15 degrees. This result outperforms related tracking works including state-of-art Fully-convolutional Network.

(2) Hand Pose Tracking Module

Two algorithms for hand pose tracking are proposed: a ROI (Region of Interests) associated particle filter-based fingertips tracking algorithm and a CNN (Convolutional Neural Network) based hand joint estimation algorithm. Two algorithms share the same input that is the result of guitar neck tracking module, and work independently in the hand pose tracking module.

For the ROI associated particle filter-based fingertips tracking algorithm, first the weighted template matching and reversed Hough Transform, which are the proposed features of fingertips are performed to the segmented hand areas in order to extract the fingertip candidates. Furthermore, a temporal grouping is applied to remove noise and group the same four fingertips (index finger, middle finger, ring finger, little finger) on the successive count maps. Then, an ROI association algorithm is utilized to associate the four fingertips with their individual trajectories on the frame-by-frame count maps. Here, for this ROI association algorithm, three patterns for tracking fingertips movement during the whole process are defined: the active pattern, adding pattern, vanishing pattern. All the tracked trajectories of fingertip candidates are fitted into these three patterns in order to solve the problem such as self-occlusion, joint-finger etc. Finally, the particle filter is utilized to track the fingertips by distributing particles within the associated ROIs of fingertips at every two adjacent frames of the video.

On the other hand, for the CNN (Convolutional Neural Network) based hand joint estimation algorithm, first three convolutional layers and two max-pooling layers output 512 channels of feature maps; then two fully-connected layers with 1024 notes respectively are connected after convolutional process; furthermore, instead of directly estimating the 3D position of each joint, a lower parameter space of a fully-connected layer with only 24 notes is utilized; 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.

Experiments are conducted using videos of guitar plays under different conditions. For the hand region segmentation, the proposed method outperforms the related works in terms of segmentation accuracy (98%) and

training efficiency (only 420 training images). For the ROI associated particle filter-based fingertips tracking algorithm, the proposed method outperforms the current state-of-art tracking algorithm with high accuracy: the mean error 6.5, 3.2, 4.9, 6.0 pixels for fore finger, middle finger ring finger and little finger respectively. For the CNN (Convolutional Neural Network) based hand joint estimation algorithm, it shows a competitive accuracy (mean error of 6.1 pixels for 16 joints) with state-of-arts but outperform them in time efficiency for both training and testing (only 4 hours for training and 0.19 ms for testing).

(3) Fingering Assessing Module

For the guitar fingering assessing, after the joints of guitarist's hand are estimated in (2), first the spatio-temporal hand pose is formulated as a two dimensional matrix for each video of guitar playing: the horizontal axis of the matrix is the 3D coordinates of 16 joints in one frame lined in one row, while the vertical axis is the frame number. Then, the proposed 3D-DCT (Discrete Cosing Transform) feature of the joints' movement of one guitar playing video is proposed by calculating the inner product between the spatio-temporal matrix of hand pose and the DCT matrix. Finally, a supervised regression SVR (Support Vector Regression) model is trained by using the 3D-DCT features to predict how well guitarist plays in the video by outputting the general score (full mark is 100 points) of the video. In another words, the system automatically evaluates the performance of guitarist based on the data-driven process without any human intervention, such as manually designing an evaluation function and etc.

Experimental results show (a) a high rank correlation (0.68) for the proposed 3D-DCT when compared with other features such as 3D-DFT (discrete Fourier transform), 3D-STIP (space-time interest points), 3D-DCT and state-of-arts, (b) better assessing result than the mid-level players, that fulfills the expectation before conducting the research: the evaluation result of the system for musical performance of the guitar playing should be better than human mid-level players.

In conclusion, in this thesis, three modules of an automatic guitar fingering assessing based on video analysis and pattern recognition are proposed: the guitar tracking module using the proposed modified RANSAC accurately tracks the guitar neck with mean tracking error of 4.2 mm and variance of 1.5 mm; the hand pose tracking module effectively tracks 2D fingertips' position by utilizing the ROI associated particle filter with mean tracking error of 5.2 pixel (7 mm), and also estimates 3D finger pose of 16 joints based on proposed CNN structure with mean estimation error of 6.1 mm; the guitar fingering assessing module evaluates guitar fingering based on 3D-DCT and SVR in a very high mean rank correlation (0.68). Generally, compared with human judge, the system predicts more accurately in guitar fingering assessing result than mid-level human player with the mean error of prediction 0.063 while the mean error of prediction of human mid-level player is 0.114 which fulfills the research goal: better assessing result than human mid-level player. Furthermore, the system also proposes some new research ideas of (1) accurately tracking a rigid object despite of translation, rotation and occlusion, which are the most difficult problems of tracking issues in computer vision; (2) gesture recognition, sign language recognition and other hand analysis problems by only doing some minor changes with the proposed hand pose estimation module; (3) data-driven method of automatically assessing the correctness of human motion.

早稲田大学　博士（工学）　学位申請　研究業績書

(List of research achievements for application of doctorate (Dr. of Engineering), Waseda University)

氏 名(Zhao, WANG)　　印()

（As of July, 2018）

種 類 別 (By Type)	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）(theme, journal name, date & year of publication, name of authors inc. yourself)
Academic Paper (International conferences)	○1. Zhao, WANG, Jun OHYA. “Tracking the Guitarist’s Fingers as Well as Recognizing Pressed Chords from a Video Sequence”. In Electronic Imaging 2016, 2016(15), pp.1-6 San Francisco, USA. ○2. Zhao, WANG, Jun OHYA. “Fingertips Tracking Algorithm for Guitarist Based on Temporal Grouping and Pattern Analysis”. In Asian Conference on Computer Vision (ACCV), pp. 212-226. Taipei, Taiwan, Oct, 2016 ○3. Zhao, WANG, Jun OHYA. “A 3D guitar fingering assessing system based on CNN-hand pose estimation and SVR-assessment”, In Electronic Imaging 2018, IRIACV-204, Burlingame, USA ○4. Zhao, WANG, Jun OHYA. “An Accurate and Robust Algorithm for Tracking Guitar Neck in 3D Based on Modified RANSAC Homography”In Electronic Imaging 2018, IRIACV-204, Burlingame, USA ○5. Zhao, WANG., and Jun OHYA. “Detecting and Tracking the Guitar Neck Towards the Actualization of a Guitar Teaching-aid System”. International conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics ICAM2015, No. 6, pp. 187-188. Tokyo, Japan, 2015 6. Gao, S., Tatematsu, N., Ohya, J., & Wang, Z.. Estimating Clean-up Robots’ Mechanical Operations of Objects Using a SLAM Based Method. In The… international conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics: ICAM: abstracts Vol. 2015, No. 6, pp. 249-250. Tokyo, Japan,2015 7. Keishi Nishikawa, Zhao Wang, Jun Ohya, Takashi Matsuzawa, Kenji Hashimoto, and Atsuo Takanishi ,“Automatic, Accurate Estimation of the position and pose of a Ladder in 3D Point Cloud”, The 5th IIEEJ International Workshop on Image Electronics and Visual Computing, 5-2C, Da Nang, Vietnam March, 2017
Academic Paper (Domestic conferences)	8. Zhao WANG, Ye LI, Jing YAN, Jun OHYA, “Study of Detecting the Frets and Strings on the Neck of the Guitar from RGBD Images towards the Actualization of an Autonomous Guitar Teaching System”, Section D7,Media Computing Conference 2014 9. Zhao WANG, Jun OHYA, “ Study of a Vision Based Method for Checking the Position of Each Finger of Guitar Players - Towards the Actualization of an Autonomous Guitar Chord Teaching System -”,Section D-11-7,IEICE Society Conference 2015. 10. Zhao WANG, Jun OHYA, “ A Method for Tracking Guitar Neck and Fingertips: Necking Tracking Robust against Occlusions Based on Geometry Analysis and Fingertips Tracking Based on Temporal Probability Map-”,FIT Information and Science Technology Forum 2015. 11. 本田 浩暉, 王 釗, 大谷 淳, 透視変換を用いたギター演奏時のネックの動画像における追跡法の検討, 画像電子学会第 280 回研究会, 03/2017 12. 前田 尚俊, 王 釗, 大谷 淳, Support Vector Regression に基づく 3 次元動画像処理による人物の動作評価法の検討, 画像電子学会第 280 回研究会, 03/2017 13. Zelin Zhang, Zhao Wang, Jun Ohya. “Hand Pose Estimation from Single Depth Images with 3D Convolutional Neural Network”, in IEICE PRMU2017-140, vol. 117, no. 391, pp. 271-276, 01/2018.