Pose Estimation for Facilitating Movement Learning from Online Videos

Atima Tharatipyakul
Singapore University of Technology
and Design
atima_tharatipyakul@mymail.sutd.edu.sg

Kenny Choo
Singapore University of Technology
and Design
kenny_choo@sutd.edu.sg

Simon T. Perrault
Singapore University of Technology
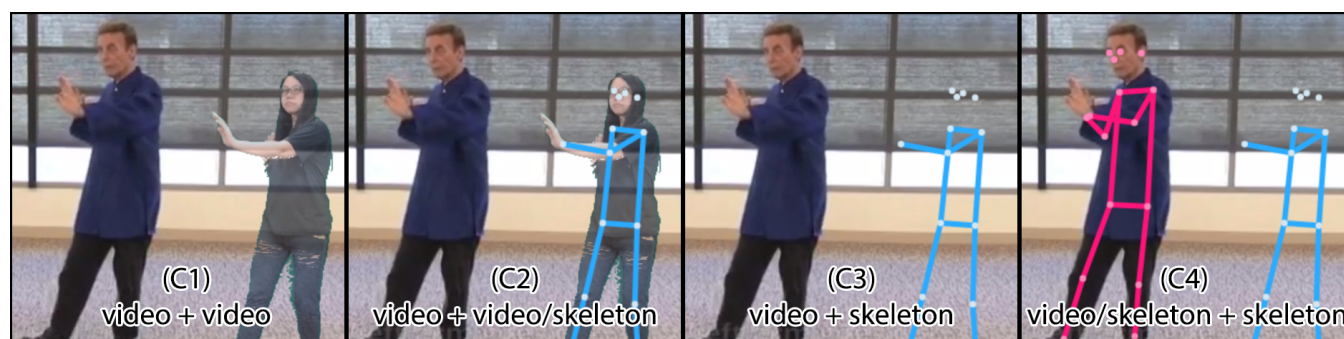and Design
perrault.simon@gmail.com

Figure 1: A study on four different types of visual feedback on a tai chi video: (C1) trainer video + user video, (C2) trainer video + user video with skeleton, (C3) trainer video + user skeleton, and (C4) trainer video with skeleton + user skeleton. The tutorial video is from https://youtu.be/ZxcNBejxlzs.

ABSTRACT

There exist a multitude of online video tutorials to teach physical movements such as exercises. Yet, users lack support to verify the accuracy of their movements when following such videos and have to rely on their own perception. To address this, we developed a web-based application that performs human pose estimation using both video inputs from the online video and web camera, then provides different types of visual feedback to a user. Our study suggests that the user's skeleton overlaid on the user's camera feed improved user performance, whereas the user's skeleton on its own or trainer's skeleton with the trainer video offered limited benefits. We believe that our application demonstrates the potential to enhance learning physical movements from online videos and provides a basis for other guidance systems to design suitable visualizations.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *User interface design.*

KEYWORDS

Movement guidance, visualization, pose estimation

1 INTRODUCTION

Online video tutorials have become a popular mean to learn physical movement. The production of the online videos, however, could range from professional editing, which considers aspects such as viewer perspectives and additional graphics, to the simplest from-the-phone recordings. The unconstrained nature of online videos makes it hard for automated processing like extracting a trainer's pose. Physical movement learning tools that require the trainer's pose to provide guidance, such as those in [17, 18, 23], rarely use online videos as a basis for training even though they are massively available and untapped resource.

The augmentation of online videos with useful feedback presents a promising mean to improve physical movement learning, as seen in previous works with the Microsoft Kinect (e.g. [2, 14]). Still, only limited number of works address different ways to represent the trainer and user (e.g. 3D model against skeleton [24]). Advances in computer vision with unconstrained camera feeds (e.g. [3, 20]) motivate us to reexamine online videos and augment them with pose information extracted from the videos. How different visualizations could facilitate movement learning from online videos remains as an open issue and is our main research question.

In this paper, we created a system that uses pose estimation to provide visual feedback to users who want to follow tutorial videos. The contributions of our study include (1) the development of a fully-working interactive system featuring configurable interface to

facilitate learning from online videos and (2) the study and insights on four types of visual feedback, relying on combinations of (a) the video feeds of both video trainers and the users and (b) pose estimation data represented as a skeleton. We used tai chi, a physical activity that requires highly accurate movements, as a case study in our controlled experiment. Our results suggest that participants were able to follow tutorials more closely using the trainer video with their own video feed that superposed with pose estimation data (skeleton).

2 BACKGROUND AND RELATED WORK

Tai Chi. Tai chi is a physical activity with slow continuous movements to promote well-being. As it requires highly accurate movements, it has been popularly used in research to help users practice it. Due to the complexity of the movements, most works employ virtual 3D teachers, either with a screen [5], head mounted displays [4, 9–12], or audio-visual-tactile devices [16]. The most similar work to ours is Stillness Moves [14], which uses the recorded weight from pressure sensing shoes and video data from the Kinect for the training. However, input from the Kinect, webcam, and other camera-based devices tend to suffer from occlusion and/or estimation error. Therefore, trainer data are usually collected in controlled settings to ensure its accuracy, as seen in [14] and other systems (e.g.[11]). While there exist works on other type of physical activity that capture data in unconstrained scenes [13], none of them address edited videos, nor compare multiple types of visual feedback as we do in this paper.

Visual Feedback Design. Sigrist et al. [17] reviewed different types of feedback strategies such as concurrent feedback, terminal feedback, fading feedback, performance-based feedback, self-controlled feedback, and bandwidth feedback. To help the user verify and improve their movement, the learning tools must provide a mean to compare the user's data with the trainer's. Hence, we can also categorize the visualization based on comparison approaches [8]: juxtaposition (putting objects side-by-side), superposition (overlaying objects on top of each other), and explicit encoding (representing relationship directly). Each approach or strategy has its own strengths and weaknesses. For instance, Timmermans et al. suggested concurrent feedback to be effective for beginners while terminal feedback may benefit more skilled users [22]. We use these insights to justify our design choices.

Video Feed vs. Skeleton Information. Despite the capabilities of the visual domain, reduced feedback visualization could prevent an overload of information for complex motor tasks [6]. One essential consideration when dealing with videos, which are considered as complex objects for comparison, is how to reduce the complexity of the videos to reduce cognitive load [7, 21]. Displaying the skeleton or skeleton on top of video feed, as seen in [2, 14] and many systems, could bring interested information to the user and reduce the cognitive load. The skeleton, however, may result in lose details and hider feedback effectiveness. For instance, Haoran et al. [24] found 3D models to be more effective than skeleton in supporting core learning. Still, none of the work mentioned above investigated the effectiveness of different visual feedback using video feed and/or skeleton information from pose estimation. Our work is thus the first study to perform such a comparison.

3 FEEDBACK AND SYSTEM DESIGN

3.1 Visual Feedback Design

We wanted our system to provide high quality feedback that would allow users to follow a tutorial video accurately. The idea was to add relevant pieces of information on the user's current pose to enable a quick side-by-side comparison.

Showing the user's body is an easy and convenient way to provide accurate feedback. After some initial testing, where the user's video feed was located too far away from the trainer, participants reported that it was hard to follow the trainer as they had to constantly look at the trainer, then their own feed sequentially. We realized that both video feeds had to be shown as closely as possible, to keep all the relevant information in the user's central vision. We decided to use background subtraction. This way, we can reduce the distance between the feeds, and the user can focus on the important parts of the video feed, i.e. their own body.

In addition to the user's video feed, we also decided to use pose estimation data. Previous work [2, 14] made use of that data by drawing a skeleton to represent the user, showing the individual position of the joints on the user's body (e.g. shoulder, legs, arms).

3.2 Implementation

We developed a web-based application that takes the URL of a tutorial video file and the user's web camera feed as inputs. The overall interface is similar to a typical video player, except that it also provides concurrent visual feedback – showing a mirror image of the user from the camera beside the trainer in the video, as demonstrated in Figure 2.
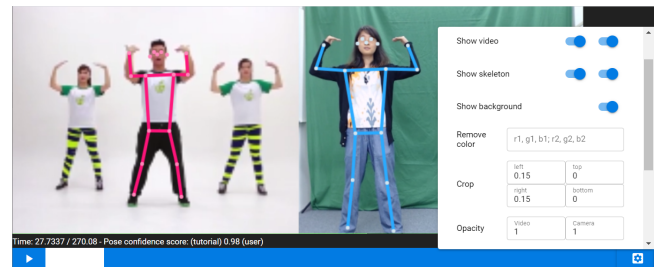

Figure 2: Screenshot of our application with options to control visibility and appearance of the videos and skeletons. The tutorial video is from https://youtu.be/3EbBHoRgn-Y.

The application supports three types of content visualization: video, filtered video, and skeleton. We support various functions such as cropping of the video frames, adjusting opacity, or removing the background (either using chroma keying with a green screen or automated person segmentation [19]). The skeleton is estimated in real-time using a state-of-the-art pose estimation library [20]. The application supports individual visibility toggling of the video feed and the skeletons of either the user or trainer.

The application facilitates comparison of the trainer and the user through natural feedback visualization. The user can compare themselves with the trainer side by side. Both feeds can be moved around, resized, and overlaid on top of another. In contrast with explicit encoding design that requires predefined pose comparison

algorithm, these juxtaposition and superposition approach leverages user's perception to realize different between their movement and the trainer's.

4 USER STUDY

We conducted a controlled experiment on each visualization type on a tai chi video to examine how each visualization type works. Our goal here was not testing all possible combination of configurations, but finding out usefulness of a trainer's skeleton, user's skeleton, and user's video feed. The four conditions we selected for our user study are: (C1) trainer video + user video (baseline), (C2) trainer video + user video with skeleton, (C3) trainer video + user skeleton, and (C4) trainer video with skeleton + user skeleton, as illustrated in Figure 1. By limiting number of conditions to four, we did not exhaust or overload participants with too many choices.

4.1 Participants

We used Latin Square to minimize the number of required participants (i.e. 4 participants for 4 conditions) and replicated it 3 times to increase the degrees of freedom for experimental error (see [1]). In total, we recruited 12 participants (7 female), aged from 22 to 37 years old ($M = 27.50$, $SD = 4.56$) for the experiment. Only one of them had rarely practiced tai chi.

4.2 Apparatus

For this experiment, we used a PC with a Xeon E5-2603 processor at 1.6 GHz, with 16 GB of RAM and a Nvidia Titan GPU with experimental software written in JavaScript running on a Chrome web browser. For the user video input, we used a Logitech C922 Pro webcam and scaled the resolution to 640×360 to match with a tutorial video retrieved from YouTube and maintain a good frame rate with the pose estimation. The computer was connected to a 27" monitor with a resolution of 2560×1440 pixels. We also used a green screen located behind the user to perform the background subtraction in real time.

4.3 Procedure

We first obtained informed consent from the participants and got them to fill in a demographics survey. They were then briefed about the experiment and continued to perform the four trials – one for each visual feedback condition (C1, C2, C3, C4). At the beginning of the trial, we asked participants to stand on a mark on the floor located 2.5 meters away from the monitor. This was to ensure that the camera captured the whole body of the participant, and the videos were within a participant's central vision. Participants were then allowed to try the condition for a period of 30 to 60 seconds for training, before proceeding with the actual trial. We used the same video[1] for the training for all conditions.

Tai chi tutorial video would then start, and the participant had to follow the movements as closely as possible. Each trial contained one move and lasted for about one minute. We chose four different moves of similar difficulty and presentation from one video[2]. The trainer repeated each move 6 times (3 times for left and right side). The original audio in the video was muted to focus participants on

[1]https://youtu.be/zDSlx3INZWk
[2]https://youtu.be/ZxcNBejxlzs

the visual feedback. After finishing a trial, participants had to fill a NASA TLX questionnaire and provide subjective feedback about the condition. After completing all four conditions, participants were asked to rank the techniques and indicated where they looked during the trials (experimenter, themselves, or both).

4.4 Design

We used a within-subject design with one independent variable: *Visual Feedback* { (C1) trainer video + user video (baseline), (C2) trainer video + user video with skeleton, (C3) trainer video + user skeleton, and (C4) trainer video with skeleton + user skeleton }. The order of presentation of the visual feedback condition was counterbalanced using a Latin Square.

We measured the angular error as a dependent variable. We computed the average angular difference of the shoulder, hip, upper and lower arms, as well as upper and lower legs between the trainer and the participant. The angular error was measured in radians and was not displayed to the participant. We excluded the head from the calculation since the participant had to tilt their head when checking the pose on-screen. We also measured the individual components of the NASA TLX as dependent variables.

Participants completed the experiment in about 60 minutes and were encouraged to take breaks between conditions. Our design is as follows: 12 participants $\times$ 4 conditions $\times$ 1 repetition = 48 trials.

4.5 Results

4.5.1 Error. We used one-way ANOVA for statistical analysis. We found a significant main effect of *Visual Feedback* on error ($F_{3,33} = 3.63$, $p = .02$). The average angular error was 0.169 radians. A Tukey HSD test for post-hoc analysis showed that the trainer video + user video with skeleton (C2) condition achieved the lowest error ($M = 0.146\ rad$), which was significantly lower than our baseline (C1, both videos, $M = 0.188\ rad$, $p < .05$). We did not find any other significant differences in terms of error, with our other conditions (C3 and C4, see Figure 1) achieving an accuracy of 0.170 and 0.174 rad respectively. The error rates are summarized in Figure 3.
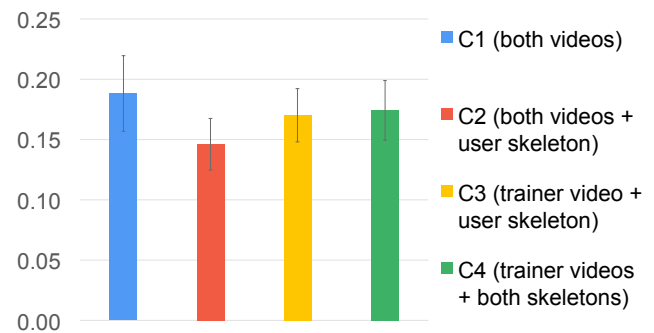

Figure 3: Average angular error in radians across conditions. Error bars show .95 confidence intervals.

4.5.2 NASA TLX. Each condition achieved an average score between 33.8 (C2) and 35.9 (C1 and C4). We did not find any significant main effect of *Visual Feedback* on the overall TLX score, nor with

any individual component, except for Physical Demand ($F_{3,33} = 3.4$, $p = .03$). Specifically, C2 ($M = 2.56$) was as deemed as significantly less physically demanding as compared to C4 ($M = 5.19$, $p < .05$). The NASA TLX scores are shown in Figure 4.

4.5.3 Ranking. In addition, C2 was ranked first or second best technique by 9 participants out of 12. We decided to sum the rank given by participants to each technique (1st = 1 point, 2nd = 2 points, etc.), with the lower the score, the better. By doing so, we found that C2 gets the lowest (best) score of 24, followed by C4 with 28 points, C1 with 33 points and C3 with 35 points.
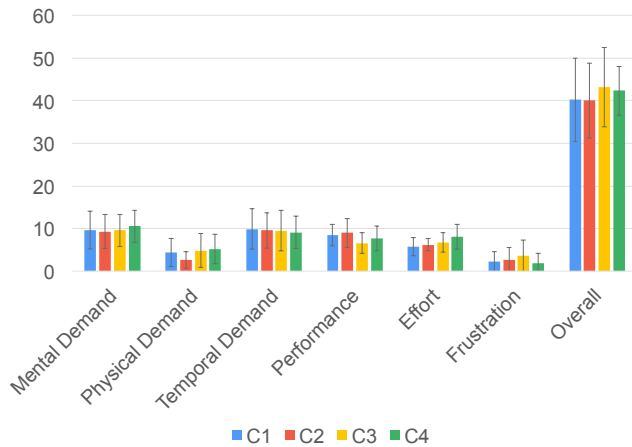

Figure 4: Average NASA-TLX for each condition. Error bars show .95 confidence intervals.

4.5.4 Subjective Feedback. Participants liked C1 (baseline) and found it "easy" (P4) and "clear to follow" (P2, P12). Participants liked being able to compare their live video with the trainer's as it helped them "correct movements" (P5, P8, P10), but was deemed as "distracting" (P7). Participants enjoyed the addition of the skeleton on themselves as it helped them focus on their own movements (P4). Participants reported that C3, where the user's video feed is not displayed, made it harder to "compare" movements with the trainer (P1, P4, P8, P12). Some participants found it simpler to use (P2, P5, P8) and helpful for catching up with the trainer (P12). Adding the skeleton on top of the trainer (C4) was deemed as helpful for comparison (P1, P9), easier to adjust their pose (P11), and overall improved the performance (P10). It, however, made some parts of the trainer hard to see (P5, P7, P14) and was mentally challenging (P5) because of too much information (P10).

Participants reported that they looked mostly at the trainer for the whole experiment (P4-P7) or few first repetitions (P1-P3, P8-P12), limiting the time experience different conditions or comparison. Finally, participants have different learning styles, and each condition could be good in different situations and/or learning goals. For instance, skeleton could help getting a better idea of angle but make hands hidden (P5, P14). P3 would like to see the trainer video with skeleton first, then the user video with skeleton. P10 suggested that the tool should be goal oriented since little error is fine for self-practice so no need for detail comparison.

5 DISCUSSION AND LIMITATIONS

Our results suggest that C2 (trainer video + user video with skeleton) allowed our participants to be significantly more accurate. It was also deemed as less physically demanding and was overall preferred by our participants. Data suggests that the user skeleton can help improving user performance. On its own, the user skeleton only offers limited benefits and makes comparison with the trainer data complicated, as shown by C3's lower performance. However, it allows the user to get a quick overview of their current pose. The addition of the user's video feed also enabled a more precise side-by-side comparison with the trainer's pose. We were originally worried that this condition would be too demanding for the participants, but the NASA TLX results suggest that it was not the case, and that the benefit of making the tutorial video easier to follow offsets any potential additional cognitive load. The skeleton on the trainer, on the other hand, could introduce visual clutter without improving user performance, as suggested by participants' comments and error of C3 versus C4. Visual clutter is not an issue in the user's case as the user already knows their own movement. Counter-intuitively, there is no need to visualize the trainer and user in the same way.

As a first step to explore automated processing of videos in the wild, we limited the case study to a video with few estimation errors. Our current system works well only when the trainer and user's whole body are viewed from the same camera angle. We also relied on an automatically estimated 2D skeleton, which may not be accurate in depicting 3D movement. We, however, believe that these limitations could be overcome in the near future by, for example, recent development in 3D human pose estimation from video [15, 24].

We found that some errors were based on a biased perception: for instance, P7 did not notice inaccuracy in their pose estimation until the experimenter pointed it out. They reasoned that since tai chi is slow, even one second of offset in terms of pose was unnoticeable when they focused on arm movement.

6 CONCLUSION AND FUTURE WORK

In this paper, we investigated different types of visual feedback to help users follow tutorial videos. Our visual feedback relied on video feeds and pose estimation data. We found out that the best combination involves the trainer's video and the user's video feed superposed with pose estimation data. That type of feedback enabled direct comparison between both video feeds and allowed users to quickly assess their current pose using the skeleton. As future work, we would like to consider other applications such as dance and explore other mechanisms to help pose correction during training.

REFERENCES

[1] Mike Allen. 2017. *The SAGE encyclopedia of communication research methods.* SAGE Publications.
[2] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface software and technology.* 311–320.
[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.
[4] Philo Tan Chua, Rebecca Crivella, Bo Daly, Ning Hu, Russ Schaaf, David Ventura, Todd Camill, Jessica Hodgins, and Randy Pausch. 2003. Training for physical

tasks in virtual environments: Tai Chi. In *IEEE Virtual Reality, 2003. Proceedings.* IEEE, 87–94.

[5] Wagner O De Morais and Nicholas Wickström. 2011. A serious computer game to assist Tai Chi training for the elderly. In *2011 IEEE 1st International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 1–8.

[6] Daniel L Eaves, Gavin Breslin, Paul Van Schaik, Emma Robinson, and Iain R Spears. 2011. The short-term effects of real-time virtual reality feedback on motor learning in dance. *Presence: Teleoperators and Virtual Environments* 20, 1 (2011), 62–77.

[7] Michael Gleicher. 2018. Considerations for Visualizing Comparison. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (1 2018), 413–423. https://doi.org/10.1109/TVCG.2017.2744199

[8] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309. https://doi.org/10.1177/1473871611416549

[9] Ping-Hsuan Han, Yang-Sheng Chen, Yilun Zhong, Han-Lei Wang, and Yi-Ping Hung. 2017. My Tai-Chi Coaches: An Augmented-learning Tool for Practicing Tai-Chi Chuan. In *Proceedings of the 8th Augmented Human International Conference (AH '17)*. ACM, New York, NY, USA, 25:1–25:4. https://doi.org/10.1145/3041164.3041194

[10] Felix Hülsmann, Jan Philip Göpfert, Barbara Hammer, Stefan Kopp, and Mario Botsch. 2018. Classification of motor errors to provide real-time feedback for sports coaching in virtual realityâĂŤA case study in squats and Tai Chi pushes. *Computers & Graphics* 76 (2018), 47–59.

[11] Takahiro Iwaanaguchi, Mikio Shinya, Satoshi Nakajima, and Michio Shiraishi. 2015. Cyber tai chi-cg-based video materials for tai chi chuan self-study. In *2015 International Conference on Cyberworlds (CW)*. IEEE, 365–368.

[12] P Kao, P Han, Y Jan, Z Yang, C Li, and Y Hung. 2019. On Learning Weight Distribution of Tai Chi Chuan Using Pressure Sensing Insoles and MR-HMD. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 1457–1464. https://doi.org/10.1109/VR.2019.8797986

[13] Rushil Khurana, Karan Ahuja, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. 2018. GymCam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–17.

[14] Han Hong Lin, Ping Hsuan Han, Kuan Yin Lu, Chia Hung Sun, Pei Yi Lee, Yao Fu Jan, Amy Ming Sui Lee, Wei Zen Sun, and Yi Ping Hung. 2018. Stillness

Moves: Exploring Body Weight-Transfer Learning in Physical Training for Tai-Chi Exercise. In *Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports (MMSports'18)*. ACM, New York, NY, USA, 21–29. https://doi.org/10.1145/3265845.3265856

[15] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7753–7762.

[16] Otniel Portillo-Rodriguez, Oscar O Sandoval-Gonzalez, Emanuele Ruffaldi, Rosario Leonardi, Carlo Alberto Avizzano, and Massimo Bergamasco. 2008. Real-time gesture recognition, evaluation and feed-forward correction of a multimodal tai-chi platform. In *International Workshop on Haptic and Audio Interaction Design*. Springer, 30–39.

[17] Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. 2013. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic Bulletin & Review* 20, 1 (2 2013), 21–53. https://doi.org/10.3758/s13423-012-0333-8

[18] Ramin Tadayon, Troy McDaniel, and Sethuraman Panchanathan. 2017. A survey of multimodal systems and techniques for motor learning. *Journal of Information Processing Systems* 13, 1 (2017), 8–25. https://doi.org/10.3745/JIPS.02.0051

[19] TensorFlow. 2019. BodyPix. https://github.com/tensorflow/tfjs-models/tree/master/body-pix

[20] TensorFlow. 2019. PoseNet. https://github.com/tensorflow/tfjs-models/tree/master/posenet

[21] Atima Tharatipyakul and Hyowon Lee. 2018. Towards a Better Video Comparison: Comparison as a Way of Browsing the Video Contents. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction - OzCHI '18*. ACM Press, New York, New York, USA. https://doi.org/10.1145/3292147.3292183

[22] Annick A A Timmermans, Henk A M Seelen, Richard D Willmann, and Herman Kingma. 2009. Technology-assisted training of arm-hand skills in stroke: concepts on reacquisition of motor control and therapist guidelines for rehabilitation technology design. *Journal of neuroengineering and rehabilitation* 6, 1 (2009), 1.

[23] Rosanna Maria Viglialoro, Sara Condino, Giuseppe Turini, Marina Carbone, Vincenzo Ferrari, and Marco Gesi. 2019. Review of the augmented reality systems for shoulder rehabilitation. *Information* 10, 5 (2019), 154.

[24] H Xie, A Watatani, and K Miyata. 2019. Visual Feedback for Core Training with 3D Human Shape and Pose. In *2019 Nicograph International (NicoInt)*. 49–56. https://doi.org/10.1109/NICOInt.2019.00017