



HAL
open science

Gesture Interaction in Virtual Reality

Cloe Huesser, Simon Schubiger, Arzu Çöltekin

► **To cite this version:**

Cloe Huesser, Simon Schubiger, Arzu Çöltekin. Gesture Interaction in Virtual Reality. 18th IFIP Conference on Human-Computer Interaction (INTERACT), Aug 2021, Bari, Italy. pp.151-160, 10.1007/978-3-030-85613-7_11 . hal-04292347

HAL Id: hal-04292347

<https://inria.hal.science/hal-04292347v1>

Submitted on 17 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

Gesture interaction in virtual reality: A low-cost machine learning system and a qualitative assessment of effectiveness of selected gestures vs. gaze and controller interaction

Cloe Huesser¹ [0000-0003-4316-4251], Simon Schubiger² [0000-0002-6723-6873], Arzu Çöltekin¹
[0000-0002-3178-3509]*

¹Institute of Interactive Technologies, University of Applied Sciences and Arts Northwestern
Switzerland, Brugg-Windisch, Switzerland

²Esri Inc., Zurich R&D, Zurich, Switzerland

cloe.huesser@fhnw.ch, sschubiger@esri.com,

*corresponding author: arzu.coltekin@fhnw.ch

Abstract. We explore gestures as interaction methods in virtual reality (VR). We detect hand and body gestures using human pose estimation based on off-the-shelf optical camera images using machine learning, and obtain reliable gesture recognition without additional sensors. We then employ an avatar to prompt users to learn and use gestures to communicate. Finally, to understand how well gestures serve as interaction methods, we compare the studied gesture-based interaction methods with baseline common interaction modalities in VR (controllers, gaze interaction) in a pilot study including usability testing.

Keywords: Gestures, VR, Interaction, Usability.

1 Introduction

The overarching goal of this project is to enable interactions in VR through gestures in affordable, efficient, and effective ways. This idea is not new, in fact the question of using gestures for human-computer interaction (HCI) is one of the persistent challenges in VR research, *e.g.*, already in 1999, Weissmann and Solomon presented a gesture recognition approach based on data gloves and neural networks [1]. Some gestures have also been examined from various human-centric perspectives, *e.g.*, object manipulation effectiveness [2], intuitiveness [3], or usability [4]. With recent popularization of VR and other extended reality (XR) devices, it became (even more) evident that much work is needed to create seamless interactions between humans and XR systems [*e.g.*, 5, 6] where standard interaction approaches do not work well [7], and speech interaction lacks privacy and feel intrusive when the user is not alone [8]. Thus, gestures have a lot of potential in VR interaction [8]. However, their development and adaptation have been hindered by several factors, *e.g.*, without dedicated devices interactions are imprecise; or they lack usability, learnability and/or intuitiveness. Here we re-examine a set of gestures, as recent developments in machine learning yielded promising results for employing off-the-shelf, ubiquitous cameras

(webcams, smartphone cameras) with decent accuracy [9]. Specifically, we select a set of gesture candidates, and use a webcam (as an addition to VR hardware) to detect and employ gestures for user interaction, and explore an avatar that can mirror the user to improve learnability of gestures. Finally, we report the results from a pilot user study to better understand the usability of our prototype implementations as well as initial subjective user experience with the gestures we implemented.

2 Methods

Gesture selection. Based on literature [*e.g.*, 10, 11] and introspection, we identified commonly used human gestures as candidates for interaction in VR. We differentiated gestures that can be tracked in the headset *vs.* using an external webcam (Table 1).

Computational approach. For the VR headset (Samsung HMD Odyssey), we used built-in solutions whereas for the webcam gesture recognition, we reviewed optical motion tracking and machine learning (ML) solutions for tracking users’ hands and head. Subsequently, we developed an ML-system to detect gestures in the live feed of a webcam (Logitech, c922 Pro Stream Webcam). We then examined human pose estimation approaches, specifically on 2D color images [14–16] and selected OpenPose [15] due to its wide distribution and the quality of documentation. Thus, we implemented our gesture tracking algorithm using OpenPose 1.5.1 for 2D images directly from the webcam. An example outcome is shown in Figure 1.

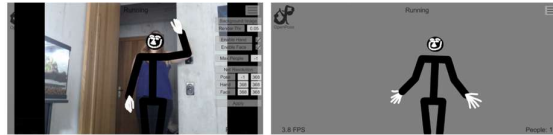
















Figure 1. OpenPose human pose detection with image (left), without image (right).

We constructed a 1750-image dataset for the gestures, and an evaluation dataset with 10 images for each gesture class for testing the ML solution. Images of each gesture from the view of OpenPose were manually classified. To minimize false predictions, we also created three *non-gesture* classes, *i.e.*, no human visible, no gesture present, user handling headset. To improve the ML performance, we augmented the training data set by flipping all images, thus doubling their number. To classify the images, we created a convolutional neural network with Keras / TensorFlow 2.0 [17]. The network consists of three Conv2D layers with max pooling and a dropout layer. In the end, the data is flattened and reduced to the number of classes with a dense layer and softmax activation [18]. OpenPose receives and processes the images, and returns a color image with the human key points, which is stored in an image folder (Figure 2).



Figure 2. Depiction of ML-pipeline.

Table 1. Overview of gesture candidates included in the study.

Name	Action	Meaning	Interactions	Tracking
a) Gestures that can be tracked with a VR headset				
nodding 	lift / lower head	yes	answer to direct yes/no question	headset sensors
head shake 	turn head left to right	no	answer to direct yes/no question	headset sensors
head tilt 	pause head tilted	thinking / not sure	show help / explanation	headset sensors
chin point 	move chin into direction of something	selecting	select object in front direction	headset sensors
wandering look 	move head / not focusing on one point	uninterested	skip intro / info / help	headset sensors
b) Gestures that can be tracked with an external webcam				
shrugging 	lift shoulders (and arms)	i don't know	repeat tutorial / intro, skip ques- tion	continuous / motion /position camera
beckoning 	move hand far to near (more than once)	nearer / bigger	zoom in, choose something	continuous / motion
stop 	move hand near to far, then stop	stop / pause	stop an action, pause	one picture / position camera
face cover 	move hand / hands before eyes	shocked, feeling uncomfortable	pause	one picture / position camera
hands up 	raise both hands open above the head	giving up	giving up	one picture / position camera
hand raise 	raise one hand open, or only the index finger stretched over head	volunteer or want to say something	question	one picture / position camera
pointing 	point with open hand or stretched index finger	indicates direc- tion	choose, go in one direction	needs testing one picture / position camera
drink 	drinking movement pretending to hold glass	drink	using resources pause	needs testing one picture / position camera
t-sign 	one hand horizontal, the other vertical underneath	time-out	pause	one picture camera

The prediction file loads in the neural network model, acquires the last pose image from OpenPose, and predicts a class for it. A server gets created at runtime and a REP socket is bound to “tcp://*:5555”. The REP socket type acts as a service for clients, receiving requests and replying to the client. When the socket receives a message from a client, it sends the classification of the last image back. Once the ML model and the pipeline was functional, we implemented an avatar that can mirror the users’ gestures using tools from Unity3D, MakeHuman, Mixamo, OpenPose, and others.

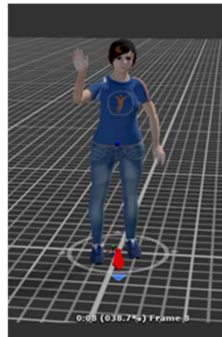


Figure 3. Waving avatar animation in Unity (right).

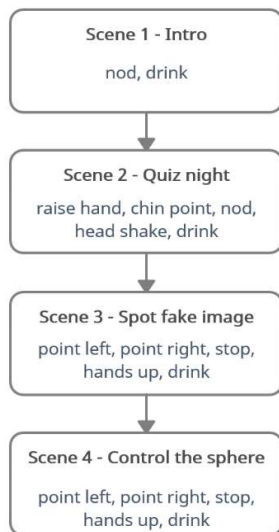


Figure 4. User study overview.

User study. We conducted a usability test to evaluate our prototypes, and have a first assessment of user experience with the gestures listed in Table 1. Our three independent variables were *interaction with gestures*, traditional VR interaction *with* and *without controllers* (i.e., gaze). Our dependent variables were *task completion rate*, *task time*, *SUS* score [12] and subjective feedback on *overall experience*, *‘naturalness’ of interactions*, *consistency of the interactions with real-world*, *perceived intuitiveness*, *satisfaction*, *effectiveness*, and *fun* (reported selectively in this paper due to page limits).

Participants. We collected qualitative feedback throughout prototype development (n=2), and recruited 6 participants (4 women, 2 men, age range 28-69) for the user study (reportedly, after six participants no new usability problems are detected [13]). We excluded participants younger than 18 for consent reasons, and ‘VR power users’ to remove possible bias.

Materials. Gesture prototypes (Table 1) and the avatar (Figure 3) were developed as described in previous sections. We then generated scenarios (Figure 4) that require the use of the selected gestures, and VR scenes to facilitate these scenarios (not shown here due to page limits) using the tools mentioned in earlier sections. We collected demographic information (e.g., age, gender, education) with a self-generated digital questionnaire, and controlled for potentially confounding variables (e.g., VR / gaming experience).

Procedure. We followed an identical protocol and strict hygiene measures with all participants (due to the covid-19 pandemic). At the experiment site, participants signed an informed consent form, filled in the demographic questionnaire, and practiced the gestures with the avatar prototype for a fixed amount of time (the avatar was included in the study to train the users for the gestures, also see Table 2, ‘Start’). Then, participants moved on to solve the same tasks three times: Once with gestures, once with controllers and once with gaze interaction (without controllers). Specific experi-

included in the study to train the users for the gestures, also see Table 2, ‘Start’). Then, participants moved on to solve the same tasks three times: Once with gestures, once with controllers and once with gaze interaction (without controllers). Specific experi-

ment tasks are shown in Table 2. We rotated the order of interaction methods to counterbalance for a possible learning effect.

Table 2. Overview tasks for the user study, including the descriptions of the tasks.

#	Task	Gesture	Gaze	Controller
Start: The avatar welcomes the participant, instructs them to act as naturally as possible, informs on the available interaction possibilities, and asks if they are ready to start.				
Task 0	Indicate that they need a break	Drink *	Gaze on pause	Click on pause with trigger
Task 1	Indicate that they are ready to start	Nod	Gaze on start button	Click on start with trigger
Quiz night (true/false): The participant is challenged with a couple of fun trivia yes/no questions. Right answers give 50 points, false give -50 points, no answer no points.				
Task 2	Select if the answer is true, false or indicate that they don't know the answer. Goal is a high score	Nod, head shake, shoulder shrug	Gaze on yes, no, or no answer button	Click on yes, no or no answer button with trigger
Fake or not (select): Participant is challenged to select two images and asked to identify which one is fake. S/he has the option to raise the hand for help/a joker. 50 points for right selection -50 for wrong, the joker reduces possible points to 25.				
Task 3	Select the fake image	Chin point	Gaze on image	Click on image with trigger
Task 4	Ask for a tip	Raise hand	Gaze on help button	Click on tip button with trigger
Task 5	Confirm selection of image	Nod, head shake	Gaze on yes or no, button	Click on yes, or no button with trigger
Command the rolling sphere: A sphere rolls from side to side over an indicated stop point. The closer the sphere stopped to the middle, the more points were won.				
Task 6	Stop the sphere as close to a certain point as possible	Stop	Gaze on stop button	Click on stop button with trigger
Task 7	Select direction in which the sphere should start again	Point direction	Gaze on direction button	Click on direction button with trigger
Task 8	Give up placement of sphere	Hands up	Gaze on finish button	Click on finish button with trigger

* We did not ask participants to perform this task in all scenes, but it is *performable* in all of them.

Once the tasks were completed, participants filled in the SUS questionnaire(s), answered numerous subjective rating questions using a 5-point Likert scale, and open-ended interview questions. SUS is a standardized questionnaire with a balanced positive and negative statements [12], and we worded the rest of the questions also as neutral as possible to avoid priming or biasing the participants.

3 Results

ML for gesture recognition. As mentioned earlier, to evaluate the ML model, we used a separate dataset (each class containing 10 images) that the model has not seen before. Figure 5 shows the model's accuracy and loss during the training over 60 epochs. Train model classified the images, and we created a confusion matrix with the

results (Figure 6). The confusion matrix shows, for example, *no_human*, *pointing_left* and *pointing_right* images were never wrongly classified, but the *t-sign* was classified as *pointing_left* in majority of the cases.

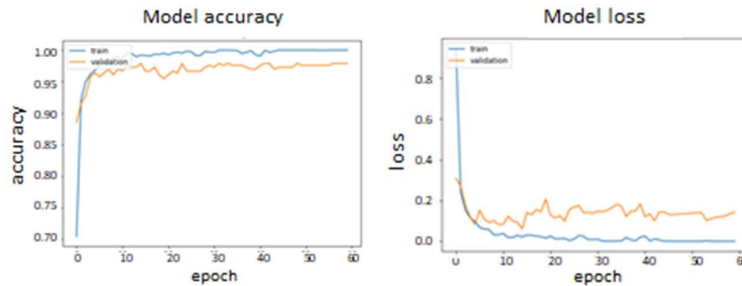


Figure 5. Model accuracy (right) and model loss (left) during training.

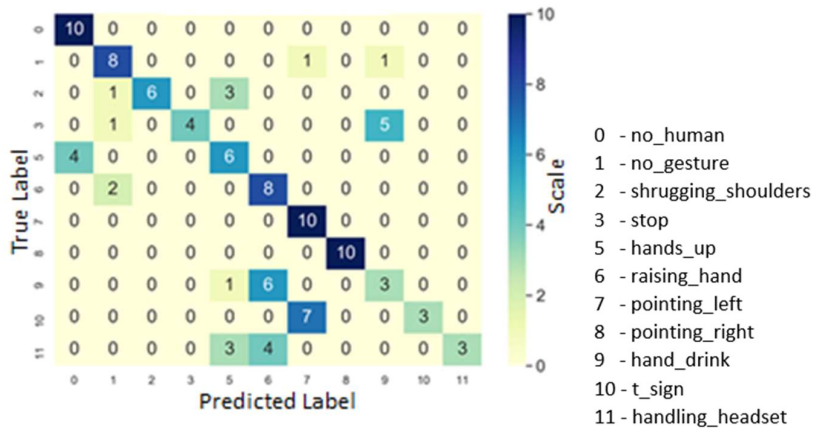


Figure 6. Confusion matrix for gesture recognition.

Usability study and pilot experiment. The overall feedback from the usability test was positive. The average SUS score is 78.75 out of 100. Since everything above 68 is considered 'usable' [12], this is good feedback. All participants successfully completed all tasks in all three conditions (controller, gaze, gesture). There was some variance in how long the users took to finish their tasks, as can be seen in Figure 7.



Figure 7. Overall average task time (left) and number of interactions (right) per interaction method.

Most of the subjective ratings in the general feedback were also positive. The participants overall liked the general concept of using gestures for interaction in several dimensions (Figure 8). One participant voted either neutral or somewhat negative in all conditions, whereas the rest of them were positive or very positive.



Figure 8. General subjective feedback from the questionnaire. Each bar shows one participant. The question is presented as a 5-point Likert scale ranging from “very negative” to “very positive”.

In a broader question (‘how would you evaluate your experience with gestures / gaze / controller?’), the interaction with the controller was overall rated better (2x very positive, 1x positive, 1x neutral) than the gestures (5x positive, 1x neutral), whereas gaze received somewhat mixed reviews (2x very positive, 3x positive, 1x negative). Questions on realism and authenticity led to overall similar results. In sum, the overall rating for gestures is quite positive (also as shown in Figure 8). At more specific level, the pointing and the hands up gestures had some negative ratings, whereas the other individual gestures were all positively rated. In the qualitative interviews, we gained some interesting insights, where participants also made positive remarks overall. However, we also observed that accidental triggering of actions frustrated the participants, and when interacting with the avatar, the delay between performing an action and seeing the action performed on an avatar seems to be irritating. On a tangential note, participants with no previous VR experience were very much fascinated by it (“wow” factor is anecdotally confirmed yet again).

4. Discussion and conclusions

We set out to examine various commonly-used gestures in human-human interactions as interaction methods in VR, a core HCI problem one can study taking computational (recognition, tracking) as well as human-centered perspectives. As a computational contribution, our ML-based gesture recognition software prototype functions reasonably well, *i.e.*, comparable to [19, 21, 22], with an off-the-shelf external webcam, when the network learns from 2D images. Gestures that are *not* about hand-tracking, such as nodding, shrugging, chin-pointing are not frequently studied, especially in combination with hand gestures as we did in this paper. While there is considerable effort in

tracking and recognizing hand gestures these are often accompanied with specialized sensors [e.g., 20]. Webcam based gesture recognition efforts, methodologically similar to ours do exist [e.g., 21, 22], but they are also often focused on hand-gestures, for example, for sign language recognition [e.g., 23]. Because tracking more parts of the body (in addition to hands) increases precision, we included head and shoulders, as well as an avatar that was mirroring the participant. While our approach works well uniquely for our specific setup where we use an external webcam to support VR interactions, comparable to the state-of-the-art systems, it clearly works better with some gestures than others (as shown in Figure 6). More work is needed to remove ambiguity from a hand or body movement that could mean more than one thing (e.g., user handling the headset vs. signaling a command) in webcam-based gesture tracking research for VR. Much like in speech recognition, we believe this will be eventually possible also with gesture recognition. A future effort that integrates the modes of gesture communication we investigated with face/emotion recognition, possibly intention recognition from gaze data might further improve the ‘vocabulary’ and range of gesture-based HCI.

As a human-centered design contribution, we tested the usability of our gesture prototypes, and comparatively assessed the initial user experience in a pilot study. Our participants received the gesture interactions overall positive / promising, similar to the earlier studies conducted on different gesture types and combinations in different VR environments [e.g., 2, 3 and 4]. Using gestures is still an experimental mode of interaction in VR. We believe more precision in tracking and recognition, further design considerations / testing, and importantly, contextualization of the gesture use, and user training are necessary. Based on the developments in these areas, gestures *can* become a main mode of interaction (such as it is in sign language), or at least a secondary supporting mode of interaction (such as it is in nonverbal body language humans constantly use). The avatar experience in this study was not precisely positive mainly due to system delays, however the approach (*i.e.*, an avatar as a virtual guide / teacher / trainer) seems very promising [24] and may lead to interesting applications such as —among many— a recent paper titled ‘Everybody Dance Now’ [25] where pose detection is used to lead / motivate / teach somebody else dance. As a tangential observation we also noted that gaze-based interaction receives mixed reactions at this point, and it would be important to keep testing this mode of interaction on its own as well a complimentary modality.

In summary, we believe gestures will remain important not only in VR but in the future of all XR interactions, and our findings suggest that usability as well as user acceptance is on its way after some more testing/training/tweaking. Our observations from the user study presented in this paper should be taken as preliminary insights due to low number of participants, and reduced trials (both of which were necessary to respect the limitations introduced by the global pandemic at the time of the study). A future controlled laboratory experiment is being planned with more participants and more trials/comparisons which will help us confirm some of the hypotheses generated based on this exploratory study.

References

1. Weissmann, J., Salomon, R.: Gesture recognition for virtual reality applications using data gloves and neural networks. In: Proceedings of the International Joint Conference on Neural Networks, Vol. 3, pp. 2043-2046, IEEE (1999).
2. Lin, W., Du, L., Harris-Adamson, C., Barr, A., Rempel, D.: Design of hand gestures for manipulating objects in virtual reality. In: Lecture Notes in Computer Science, (pp. 584-592). Springer, Cham (2017).
3. Frey, G., Jurkschat, A., Korkut, S., Lutz, J., Dornberger, R.: Intuitive hand gestures for the interaction with information visualizations in virtual reality. In: Tom Dieck, M.C. and Jung, T. (eds.) *Augmented Reality and Virtual Reality: The Power of AR and VR for Business*. pp. 261–273. Springer International Publishing, Cham (2019).
4. Cabral, M.C., Morimoto, C.H., Zuffo, M.K.: On the usability of gesture interfaces in virtual reality environments. In: Proceedings of the 2005 Latin American conference on Human-computer interaction - CLIHC '05. pp. 100–108. ACM Press, New York, New York, USA (2005).
5. Çöltekin, A., Griffin, A.L., Slingsby, A., Robinson, A.C., Christophe, S., Rautenbach, V., Chen, M., Pettit, C., Klippel, A.: Geospatial information visualization and extended reality displays. In: *Manual of Digital Earth*. pp. 229–277. Springer Singapore, Singapore (2020).
6. Çöltekin, A., Lochhead, I., Madden, M., Christophe, S., Devaux, A., Pettit, C., Lock, O., Shukla, S., Herman, L., Stachoň, Z. and Kubiček, P.: Extended reality in spatial sciences: a review of research challenges and future directions. *ISPRS International Journal of Geo-Information*, 9(7), p.439 (2020).
7. Çöltekin, A., Hempel, J., Brychtova, A., Giannopoulos, I., Stellmach, S., Dachsel, R.: Gaze and feet as additional input modalities for interaction with geospatial interfaces. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* III-2, 113–120 (2016).
8. Maloney, D., Freeman, G., Wohn, D.Y.: Talking without a Voice. In: Proceedings of the ACM Human-Computer Interaction 4, 1–25 (2020).
9. Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shajaeizadeh, M., Guo, L., Kohlhoff, K., Navalpakkam, V.: Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications* 11, 4553 (2020).
10. Morris, D.: *Bodytalk: A World Guide to Gestures*. Jonathan Cape (1994).
11. Pease, A.: *Body language: How to read others' thoughts by their gestures*. London: Sheldon Press (1988).
12. Brooke, J.: SUS -A quick and dirty usability scale Usability and context. In: Jordan, P.W., Thomas, B., McClelland, I.L., and Weerdmeester, B. (eds.) *Usability evaluation in industry*. pp. 189–196 (1996).
13. Nielsen, J.: How Many Test Users in a Usability Study? <https://www.nngroup.com/articles/how-many-test-users/>, last accessed 2021/06/10.
14. Oved, D.: Real-time human pose estimation in the browser with TensorFlow.js | by TensorFlow | TensorFlow | Medium, <https://medium.com/tensorflow/real-time-human-pose-estimation-in-the-browser-with-tensorflow-js-7dd0bc881cd5>, last accessed 2021/06/10.
15. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions of Pattern Analysis and Machine Intelligence*. 43, 172–186 (2018).
16. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: BlazePose: On-device real-time body pose tracking. *arXiv*. (2020).

17. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016 (2016).
18. Shanmugamani, R.: Deep learning for computer vision: Expert techniques to train advanced neural networks using TensorFlow and Keras. Birmingham Mumbai: Packt (2018).
19. Ahlawat, Savita, Vaibhav Batra, Snehashish Banerjee, Joydeep Saha, and Aman K. Garg.: Hand gesture recognition using convolutional neural network. In International Conference on Innovative Computing and Communications, pp. 179-186. Springer, Singapore (2019).
20. Hayashi, E., Lien, J., Gillian, N., Giusti, L., Weber, D., Yamanaka, J., ... & Poupayev, I.: RadarNet: Efficient gesture recognition technique utilizing a miniature radar sensor. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-14) (2021).
21. Golash, R., & Jain, Y. K.: Trajectory-based cognitive recognition of dynamic hand gestures from webcam videos. *Hand*, 6, 7. International Journal of Engineering Research and Technology. ISSN 0974-3154, Vol. 13, Number 6, pp. 1432-1440, (2020).
22. Agrawal, M., Ainapure, R., Agrawal, S., Bhosale, S., & Desai, S.: Models for hand gesture recognition using deep learning. In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), pp. 589-594, IEEE (2020).
23. Bendarkar, D., Somase, P., Rebari, P., Paturkar, R., & Khan, A.: Web based recognition and translation of American sign language with CNN and RNN. International Association of Online Engineering, <https://www.learntechlib.org/p/218958/>, (2021).
24. Khan, O., Ahmed, I., Cottingham, J., Rahhal, M., Arvanitis, T.N. and Elliott, M.T.: Timing and correction of stepping movements with a virtual reality avatar. *PlosOne*, 15(2), p.e0229641 (2020).
25. Chan, C., Ginosar, S., Zhou, T., Efros, A.: Everybody dance now. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), Oct., pp. 5932-5941, (2019).