STAG: Smart Tools and Applications in Graphics (2021)
P. Frosini, D. Giorgi, S. Melzi, and E. Rodolà (Editors)

Exploring Upper Limb Segmentation with Deep Learning for Augmented Virtuality

M. Gruosso[1], N. Capece[1], and U. Erra[1]

[1]University of Basilicata, Department of Mathematics, Computer Science, and Economics – Potenza, Italy

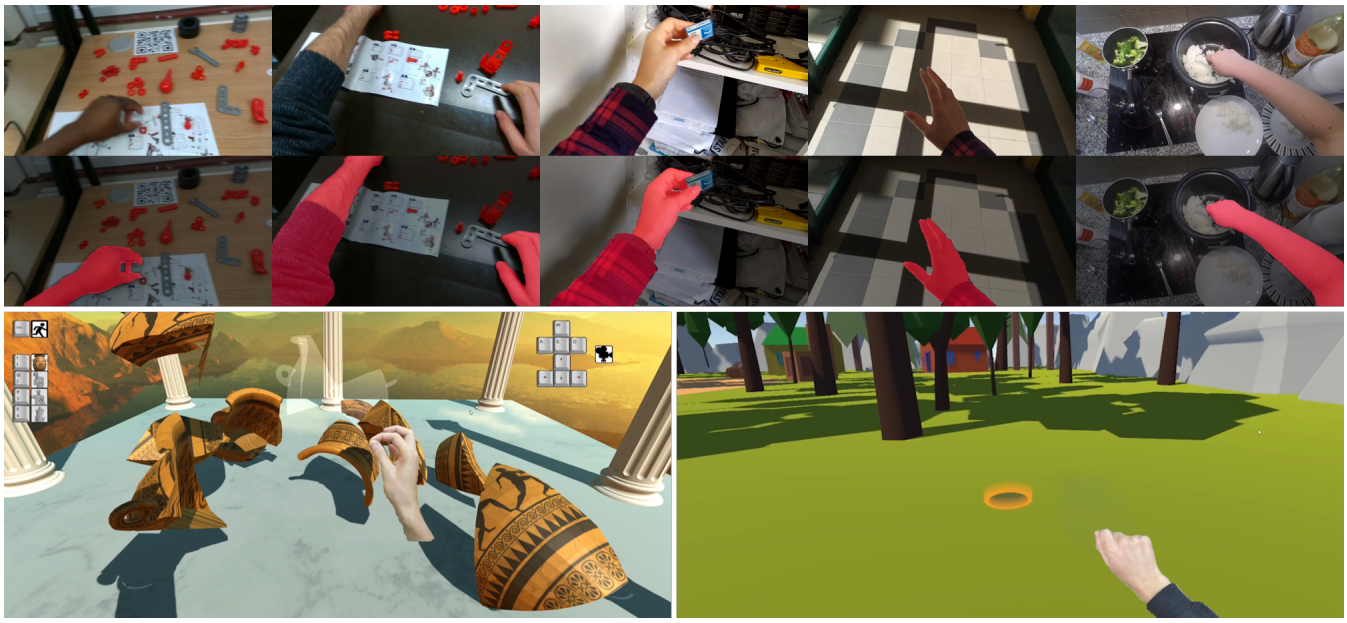

Figure 1: *We propose a deep learning-based approach to enhance the user's sense of presence in virtual environments (VEs) allowing users to see their upper limbs instead of virtual hands (bottom panel). Hand and arms are captured using a common RGB camera positioned on the Head Mounted Display (HMD) or user's head. Then, images are processed by our upper limb segmentation network, which proved to be robust to different skin tones, lighting conditions, clothes, and occlusions (top panel). Finally, the segmented human limbs are visualized into the VE, while the interaction is allowed via a Leap Motion controller.*

Abstract

Sense of presence, immersion, and body ownership are among the main challenges concerning Virtual Reality (VR) and freehand-based interaction methods. Through specific hand tracking devices, freehand-based methods can allow users to use their hands for VE interaction. To visualize and make easy the freehand methods, recent approaches take advantage of 3D meshes to represent the user's hands in VE. However, this can reduce user immersion due to their unnatural correspondence with the real hands. We propose an augmented virtuality (AV) pipeline allows users to visualize their limbs in VE to overcome this limit. In particular, they were captured by a single monocular RGB camera placed in an egocentric perspective, segmented using a deep convolutional neural network (CNN), and streamed in the VE. In addition, hands were tracked through a Leap Motion controller to allow user interaction. We introduced two case studies as a preliminary investigation for this approach. Finally, both quantitative and qualitative evaluations of the CNN results were provided and highlighted the effectiveness of the proposed CNN achieving remarkable results in several real-life unconstrained scenarios.

CCS Concepts

• ***Computing methodologies*** → ***Virtual reality; Mixed / augmented reality; Image segmentation; Neural networks;*** *Perception; Image processing;*

DOI: 10.2312/stag.20211483

1. Introduction

With the evolution of computer graphics and VR, 3D scenes have been developed with an ever-increasing level of realism in recent years. Although HMDs have allowed the user to view the virtual scene, they do not guarantee that the user will feel involved in the virtual experience. One way to increase the user immersion is to provide multi-sensory feedback and a virtual body (avatar), which is mapped to the user's movements [MBS*11]. In particular, 3D mesh and virtual representations of hands or limbs are employed when the user is personally involved and observes the scene and his body from a first-person point of view (egocentric vision) [BLAL19]. In this context, one of the main research questions concerns the sense of embodiment, which includes the sense of self-location (related to the spatial experience of being inside a body and the relation between one user and his own body), body ownership (associated with the visual body appearance) and agency (related to the body control) [KGS12; AHTL16]. Moreover, the user experience in VEs is strongly influenced by the sense of presence. Different from self-location, presence concerns the sense of being inside the VE (tele-presence or personal presence) that is responsive to the human presence (environmental presence) and allows interacting with others in the same virtual place (social presence) [Hee92]. There are several levels of immersion and presence obtained by combining different technologies and devices that connect the real world to the virtual world, producing a reality-virtuality continuum [MTUK95]. The transition between the real and the purely virtual is characterized by real environments that include virtual components (Augmented Reality, AR) or, as an opposite concept, virtual environments that integrate real elements (Augmented Virtuality, AV), resulting in mixed reality (MR) [BOJ*09]. For example, 3D scanning and photogrammetry are often used to populate virtual scenes with real-world information [GȘ18; PPC21]. These techniques have a cost in terms of computational resources and can not be easily implemented in real-time [MTRW17]. On the other hand, existing low-cost AV approaches provided stereoscopic video-see-through systems merging real egocentric views and virtual scenes. They usually employed traditional techniques to segment the user's body and subtract the original background, such as chroma-key [BSRH09; GFG15] and depth-based methods [LCBL16; RAS19], which are limited to constrained situations. Recently, deep learning preliminary solutions were proposed to overcome traditional segmentation methods, showing promising results [GPK*18; PD19; GPT*20].

In this paper, we propose an AV pipeline to increase the user immersion, presence and embodiment in VEs through the segmentation of the user's upper limbs. To achieve our goal, we used a CNN to segment users' upper limbs captured by an RGB common camera from an egocentric point of view and in real-life unconstrained scenarios. We collected a large-scale dataset consisting of real-life well-labeled images and used those data to train our segmentation neural network. The trained CNN achieved interesting and accurate results, as shown in the top panel of Figure 1. The first row shows some input images, while the second one depicts the predicted masks overlapped on each input. In particular, the first two images are from the Meccano dataset [RFLF21], the third and fourth were captured in challenging indoor scenes, and the last belongs to the Epic-Kitchen dataset [DDF*18]. Inspired by the results obtained, we defined a well-structured pipeline (see Section 4) considering the Leap Motion controller for hand tracking and our CNN to segment images captured by a simple RGB monocular camera placed on the user's head or HMD. Then, the CNN output masks extracted human limbs, which were streamed in the VE using a 2D sprite. To implement and test our pipeline, we prepared two case studies by exploiting existing applications with two different hardware configurations, as reported in the bottom panel of Figure 1 (Archaeo Puzzle [CEGA20] and Freehand-Steering [CCE*20] applications, respectively). To interact with Archaeo Puzzle, we used the Leap Motion placed on the desk in front of the user, as indicated by the authors. In addition, an egocentric RGB camera was placed on the user's head with a band, and the 3D scene was visualized with a monitor. Instead, in the case of the Freehand-Steering application, we mounted the Leap Motion controller and the RGB camera on the HMD. Using the Leap Motion controller, we faced another AV challenge, which involves tracking real-world objects and matching their positions in the VE to provide users with realistic interaction feedback [NAA18]. Furthermore, we provided both quantitative analysis and qualitative assessment to prove the effectiveness of the proposed segmentation approach and validate our deep learning method (see Section 6).

The remainder of this paper is structured as follow: in Section 2 an overview of the related work was provided; Section 3 shows the proposed CNN an datasets details; Section 4 describes in details our AV pipeline; Section 5 describes the chosen case studies, in which we included our AV pipeline; Section 6 discusses our segmentation CNN results and, finally, the Section 7 summarizes the conclusion, limitations, and future research directions.

2. Related Work

In this Section, we provide an overview of the most important AV approaches (Section 2.1), highlighting both traditional and deep learning-based methods for hand and arm segmentation (Section 2.2). Only methods based on RGB camera input were considered since they are strongly related to the proposed pipeline.

2.1. Augmented Virtuality

The growing development of technologies based on AV has allowed increasing the level of realism of VEs and better engage users in the VR experience [PPC21], using HMDs from an egocentric vision or visualizing through a monitor from a third-person perspective [BLAL19]. One of the first AV application was developed by Regenbrecht *et al.* [ROW*03; RLK*04]. They designed a prototype AV system for remote collaboration in an industrial context, called "cAR/PE!", which allowed participants to see real-world in a video-mediated way, share presentations, visualize 3D model and manipulate them. Moreover, AV can be used to add real historical artifacts into cultural heritage projects enhancing the educational impact. In that context, Gheorghiu *et al.* [GȘ18] built a VR application using Unity3D and added human characters with historical clothes and real video hotspots placed as AV content in the virtual scene. AV is also widely used in other fields, such as stress simulations and hazard recognition. Neges *et al.* [NAA18] proposed an AV approach based on a complex hydraulic system [AWAN17] to

simulate a stress scenario and address it in a controlled and safe VE instead of reality. Similarly, Bhandari *et al.* [BHB*20] presented an AV scenario based on construction sites for workers' safety training that can help them to identify risks and improve their safety decision process. The AV component consisted of real static images or videos added to the VE.

Contrary to the previous works, we augmented the VEs integrating real human limbs as a dynamic and interactive component of the 3D scene.

2.2. Hand and Arm Segmentation

The main challenge is to enhance the users' sense of presence and embodiment in VE, allowing them to see their hands and arms instead of a virtual representation during the VR interaction. One of the first works was proposed by Bruder *et al.* [BSRH09]. They developed a skin detection algorithm to segment user's limbs and displayed them in the rendered virtual view. This approach is obviously bound to a certain complexion, and many errors can occur due to color similarities with parts of the background. Additionally, they did not consider clothed arms. To solve some of these problems, a green-screen setup can be used. Indeed, McGill *et al.* [MBMB15] captured users' when typing on computer keyboards in a green box and extracted the foreground information using a traditional chroma-key technique, i.e., the HSV thresholding. Traditional methods based on green-screen and, in general, color discrimination have some intrinsic limitations due to the need for specific scenarios and lighting/color conditions. To overcome those limitations, deep learning-based approaches were designed. Preliminary solutions based on a convolutional neural network for image segmentation were proposed in [GPK*18; PD19]. In particular, Gonzalez *et al.* [GPK*18] trained a fully convolutional network (FCN) [LSD15] with a semi-synthetic dataset obtained through a chroma-key technique. They only reported few qualitative results. Pigny *et al.* [PD19] proposed a U-Net model [RFB15], which was trained using images captured from both an egocentric and a third-person viewpoint. Also, in this case, egocentric training images were obtained using chroma-key techniques. Their model reported some segmentation errors but showed promising results for AV applications. Recently, Gonzalez *et al.* [GPT*20] faced the hand and arm segmentation problem designing a deep neural network model extending their previous work. They trained a CNN using their own semi-synthetic egocentric arm segmentation dataset. Although their model achieved very interesting results, several false positive and segmentation errors in color similarities between background and foreground were obtained.

To the best of our knowledge, the proposed work is the first to design a deep learning approach for upper limb segmentation from an egocentric vision, achieve remarkable segmentation mask accuracy in unconstrained and challenging scenarios, and integrate this CNN in a well-designed AV pipeline.

3. CNN for Augmented Virtuality

We designed a deep learning-based approach to automatically segment human hands and arms from an egocentric point of view. It consists of a CNN for egocentric upper limb segmentation in unconstrained real-life environments. It achieved impressive results and was robust to a great variety of scenarios, *e.g.*, different skin tones, clothes, lighting conditions, dynamic user/camera movements, and occlusions. Our CNN is based on the state-of-the-art DeepLabv3+ model [CZP*18] and is characterized by the encoder-decoder architecture. The encoder extracts low-level features and semantic information from the input image, while the decoder provides segmentation masks recovering spatial and detailed object boundary information. The encoder includes three modules. The first one is a backbone network based on the Xception-65 model [Cho17] adapted by Chen *et al.* [CZP*18] to the task of semantic segmentation to improve the performance with faster computation compared to other models. In detail, depthwise separable convolution replaced max pooling operations of the original Xception network and further batch-normalization and ReLU were added. Then, it is followed by the atrous spatial pyramid pooling (ASPP) [CPK*17] and a 1×1 convolutional layer. The ASPP module allows to obtain a better segmentation capturing multi-scale context information and consists of three atrous convolutions, a 1×1 convolution, and an image pooling layer in parallel. Atrous convolutions [PKS15] can provide a larger field of view and take more context into account, but without increasing the computational cost and number of parameters. Moreover, depthwise separable convolutions were used in the ASPP module and the combination with atrous convolutions is particularly suitable for real-time semantic segmentation [LK19b]. The features extracted through the backbone network and the encoder output are then passed to the decoder, built using convolutional and bilinear upsampling operations.

One of the main challenges in the hand and arm segmentation task is the lack of large datasets with accurate annotations. Synthetic or semi-synthetic images are usually collected since obtaining labels is easy and low-cost. They are often taken in VEs or through a constrained green screen setup. For this reason, the images often look unrealistic or artificial and could lead to poor results in real-life use cases. Although some datasets containing real-world RGB photos are available, only hands up to the wrist and bare arms are labeled or contain few data, low-quality images, or coarse ground truth masks. Therefore, we collected a large-scale well-annotated upper limb segmentation dataset for CNN training that overcomes the limitation of existing datasets. It includes about 46 thousand images in egocentric vision from two subsets. The first contains the best data from the EDSH [LK13] and TEgO [LK19a] datasets, which show indoor and outdoor scenarios with different lights, male users' skin tone, and occlusions caused by objects. In particular, we manually inspected all images and labels and discarded incorrect or partially annotated data. The second subset is our EgoCam dataset, whose images show inter-hand occlusions, male and female subjects with various clothes captured in real-life indoor/outdoor and simple/cluttered scenes. Those images were manually annotated. Since all collected data had a different resolution, aspect ratio and orientation, we performed square crop and spatial resize to 360×360 as a pre-processing phase to align data dimensions and accelerate training. The upper limb segmentation dataset was then divided into 43.837 images for the training set and 2.184 for the test. It is available for research purposes [21].

The CNN training was performed using network weights

pre-trained on ImageNet [RDS*15] and MS-COCO [LMB*14] datasets, which are publicly available on the DeepLab project page [19]. We performed several experiments by modifying the training hyperparameters and model configuration. The best performances were obtained following the training protocol suggested by Chen *et al.* [CPSA17], and setting the base learning rate to 0.0001 and batch size to 8. In detail, we used cross-entropy as the loss function and the stochastic gradient descent optimization algorithm with momentum (SGDM) and polynomial learning rate policy. Moreover, GPU acceleration through one Nvidia Titan Xp GPU with 12GB memory was used. Finally, data augmentation with random left/right flip was applied during the training phase to avoid network overfitting. Our CNN was trained for 90K iterations until convergence (about 16 epochs) and achieved remarkable results in several real-life scenarios.

4. Pipeline

We designed a robust AV pipeline that considers the hardware device configurations, the upper limb segmentation using a well-designed CNN approach, and the implementation of our AV method applied to two virtual scenes: the first is a VR scene developed from Caggianese *et al.* [CCE*20], and the second is a desktop VE scene developed from Capece *et al.* [CEGA20]. As reported in Figure 2, our hardware configuration is based on HTC Vive HMD, the Leap Motion controller as hand tracking device, and a single RGB camera to take the upper limb pictures mounted in egocentric mode.

During the VR experience, the user's hands were tracked from the Leap Motion, and their movements were projected in the VE. In this way, the user can interact with the virtual objects via a simple freehand approach. To allow the user's hands interaction with virtual objects, we used the hand 3D meshes without the renderer component keeping enabled their mesh colliders. We placed a 2D sprite in front of the main camera component and the segmented real upper limb was displayed on it. To keep the 2D sprite always in the same camera field-of-view (FOV), we added such component as a child in the camera component hierarchy.

Our CNN continuously processes the frames captured by the RGB camera device placed on the HMD or on the user's forehead through a band in an egocentric mode. The CNN output is the upper limb binary segmentation mask with 0 for background and 1 in the case of foreground pixels. The binary mask is then used in a further image processing step, which subtracts the background from the input RGB frame by removing the image portion without the upper limb. In this way, the processed frame consists only of the upper limb in foreground that is streamed in the VE through the 2D sprite component. Whether no human limbs are captured by the camera, then the output mask contains only 0 values and the 2D sprite is transparent.

Our pipeline was implemented in the applications used as case studies (see Section 5). Such applications were developed using Unity 3D as game engine exploiting the C# programming language together with Steam VR SDK (version 2.0). The pipeline was tested on a workstation with an Intel Core i7-3rd generation CPU, 16GB RAM, and one Nvidia Titan Xp GPU with 12GB memory. This hardware features are enough to execute the applications and the CNN inferences in real-time. We used the HTC Vive (version 1) for the VR scene developed from Caggianese *et al.* [CCE*20].

5. Case Studies

Our AV approach was tested using two existing applications as case studies. The first application, called Archaeo Puzzle, represents a non-immersive VE in terms of visualization but immersive in terms of interaction. The second one is an immersive VE freehand-steering locomotion application in terms of both visualization and interaction. A video demo is available on the project web page [21].

5.1. Archaeo Puzzle

Archeo Puzzle is a desktop application that considers the Leap Motion controller placed on a desk in front of the user position. The proposed system allows the 3D reconstruction of historical artifacts, split into multiple pieces scattered around the scenes. The original application provided a 3D meshes representation as visual feedback of the hands. We introduced our pipeline in this case study by shows the real users' limbs in the VE, as reported in the Figure 3.

Although the user immersion feeling and sense of presence are limited due to the non-use of the HMD device, we incremented them by displaying the real user's upper limbs. The main advantages are the enhancement of user engagement during the virtual experience and, more importantly, the reduction of negative immersion feeling feedback. Those distracting feedback are usually caused by tracking artifacts, which are due to hand joints tracking errors, as shown in Figure 4. Indeed, Leap Motion hand tracking suffers from interference issues caused by lights directed towards its infrared sensors, producing tracking loss. In those cases, the corresponding 3D virtual hands acquire an unnatural hand pose, causing user disorientation and loss of immersion.

Figure 3 shows a typical Archaeo Puzzle scenario, in which the user has to take each piece and position it on the 3D mesh ghost of the historical artifact, which is obtained disabling its renderer component which steer the user to reconstruct the 3D puzzle easily. Although 3D virtual hands can help users to better interact with the 3D pieces, they are artificial and different in terms of dimension and naturalness of movements [SLD*18]. For this reason, the sense of presence and ownership is strongly increased thanks to our AV technique. This is highlighted especially for the pinch gesture that we used to grab the objects, in which the index fingertip touches the thumb fingertip. Indeed, the user can observe the exactly performed movement of his hands even in the VE, through the segmented image of his upper limbs.

5.2. Freehand-Steering Application

Freehand-Steering is a VR immersive system developed to compare four spatial steering locomotion methods, Palm, Index, Gaze and HMD Controller. It was inspired by Mine's seminal work [Min95]. We considered only the Palm and Index steering methods to implement our AV pipeline. We do not consider the Gaze-steering way

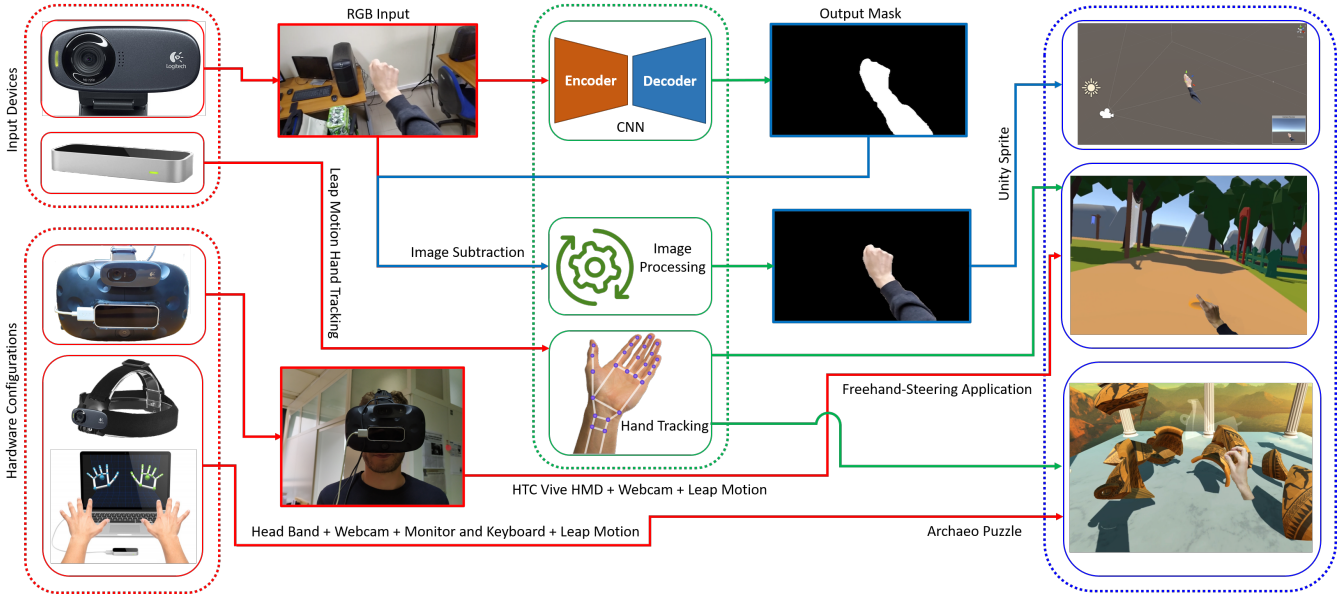

Figure 2: *Our proposed pipeline: the top red block shows the input devices used to track the upper limb and capture the correspondent RGB images in real-time; the bottom red block shows our two hardware configurations and each of them is connected with its own case study application; the green block represents the input processing core, showing our encoder-decoder CNN, the image processing step used for background subtraction and upper limb extraction, and the synchronized hand tracking phase; the last blue block contains the two case studies with the real user's hand and, at the top, the Unity 3D scene showing the upper limb streamed to a dynamic 2D sprite at runtime.*

Figure 3: *Archaeo Puzzle with AV. Real human upper limbs segmented with our CNN are placed in the camera FOV simulating the 3D virtual hand position defined in the original project. The software behaviors are keeping unchanged.*

Figure 4: *External directed light interference error on the Leap Motion causes tracking artifact and consequently 3D mesh screwing. This scenario provides a bad user immersion feeling feedback.*

because only the direction selection is defined using the gaze direction based on the HMD position and orientation, but the start and stop movements were provided using the open and closed hand gestures like Palm-steering method. The considered methods use gesture recognition through a single hand, tracked using the Leap Motion controller placed in an egocentric position on the HMD. Figure 5 shows an example of the use of Palm-steering technique with our AV.

Our application foresees continuously controlled movements by keeping the user hand visible in the Leap Motion FOV and the constant locomotion speed a priori defined. If the hand comes out of the Leap Motion FOV, then the traveling stops.

Palm-steering method allows users to define their travelling direction using their palm orientation and, in particular, the tracked palm outgoing vector. A yellow placeholder is visualized on the ground indicating the position to reach. To start the locomotion, the grab gesture is used from the only tracked hand used to define the travelling direction. Instead, the open hand gesture was used to stop the locomotion. Although Palm-steering method allows the user to move in all directions without rotating its heads, the immersion feeling is decreased due to the use of a 3D virtual hand representation. Also, in this case, we replaced the virtual counterpart

Figure 5: *Freehand-Steering application with AV. The image shows an example of Palm-steering technique discussed from Caggianese et al. [CCE*20] but using the real user's hand and arm in VR.*

with our CNN segmented upper limb, which can provide positive visual feedback to the user. Moreover, since the Leap Motion controller follows the HMD on which it is mounted, the probability of light interferences are increased as the risk of getting the same screwing on the 3D virtual hand, as discussed in Section 5.1.

Index-steering method allows users to define their control travelling with only the index finger raised, as shown in Figure 6. Using their index fingertip, the users have to point out the position they want to reach, indicated with a yellow placeholder.

Figure 6: *An example of Index-steering method with our proposed AV. As can be seen, the real segmented user's upper limb replaces the virtual hand, increasing the user immersion feeling and ownership.*

The pinch gesture is adopted to start and stop locomotion. In particular, users have to close their middle fingertip toward their thumb fingertip to start the locomotion. On the other hand, to stop the locomotion, the users have to reopen these fingers. Also, in this case, we noticed that the use of AV allows the user to feel fully engaged in the VE, increasing the sense of presence and embodiment. Furthermore, the AV solution reduced the bad visualization feedback due to the Leap Motion hand tracking errors.

6. Analysis of the CNN Results

In this Section, we reported a well-structured analysis of the results obtained with our segmentation CNN for the proposed AV pipeline. We evaluated the effectiveness of the proposed segmentation approach testing the CNN with our upper limb segmentation test set. In particular, we computed standard metrics for the segmentation task [MBP*21; GCE21], *i.e.*, Accuracy (Acc), Intersection over Union (IoU), and mean F1 score. Other metrics often used are precision and recall, but we did not consider them separately since the F1 score is the harmonic mean of the two. As reported in Table 1, our CNN obtained values greater than 97% for the metrics calculated on the overall test set. In this case, the average values of the first two metrics (mAcc and mIoU) computed over the total number of classes are usually considered since Acc and IoU may not be reliable, for example, due to unbalanced classes [LR19]. The second and the third metric group shows the per-class metrics. The worst values were obtained for the IoU and mF1 score, which are more sensitive to errors on the boundary of objects to be segmented. Instead, the Acc value takes into account only the pixel classification accuracy.

Moreover, we visually inspected the segmentation masks predicted by the CNN. Some examples are shown in Fig. 7. A good level of accuracy was obtained in different situations, as can be noted by comparing our predictions (second row) with the ground-truth (GT) labels (last row). In particular, the first four images show male hands illuminated by ambient light with a simple background, the fifth and sixth were captured in cluttered scenes with a flashlight, while the seventh image illustrates male hand and arm with artificial indoor light and positioned on a cluttered desk. Instead, the last three panels are video frames captured during user/camera movement and show a female upper limb against the light. We noticed that our approach was also robust to a little amount of motion blur.

Furthermore, the network achieved good performance also considering the computation time, reaching an average inference time of 0.02 seconds per image (see Section 4 for details on the workstation used). In this way, real-time performance of our AV pipeline is ensured, avoiding the lack of upper limb frames in the VE. Hence, the obtained results highlighted the effectiveness of the proposed method and the robustness of our CNN under several real-life unconstrained conditions.

7. Conclusion

This work shows an AV pipeline based on upper limb segmentation via a well-designed CNN-based approach and well-structured free-hand methods to interact and/or locomote in the VEs. We integrated our AV pipeline in two existing applications, which we considered as case studies. The first application, called Archaeo Puzzle, allow users to interact with a VE through their tracked hands and reconstruct historical artifacts that are split into several pieces like a 3D puzzle. As explained in Section 5.1, the user can interact and grab the 3D pieces through the pinch gesture. The second application, the freehand-steering application, allows users' locomotion in the VE through one tracked hand and well-designed gestures that define specific locomotion methods. As explained in Section 5.2, we consider the Palm and Index methods. Our AV pipeline's main challenge is to increase the user immersion feeling, embodiment, and sense of presence during the VR experiences. We address this challenge by displaying the segmented user upper limbs in the VE with a dynamic 2D sprite. Furthermore, we explained in detail our CNN, which is structured as an encoder-decoder architecture based on

Overall			Limb			Background		
mAcc	mIoU	mF1	Acc	IoU	mF1	Acc	IoU	mF1
98.84	97.69	97.35	97.96	95.93	96.31	99.72	99.46	98.39

Table 1: *The metric values in percentage related to the overall test set and for each class are reported. In particular, Intersection over Union (IoU), Accuracy (Acc), and mean F1 score (mF1) are considered. In the case of the whole test set, the mean values of IoU and Acc are computed.*

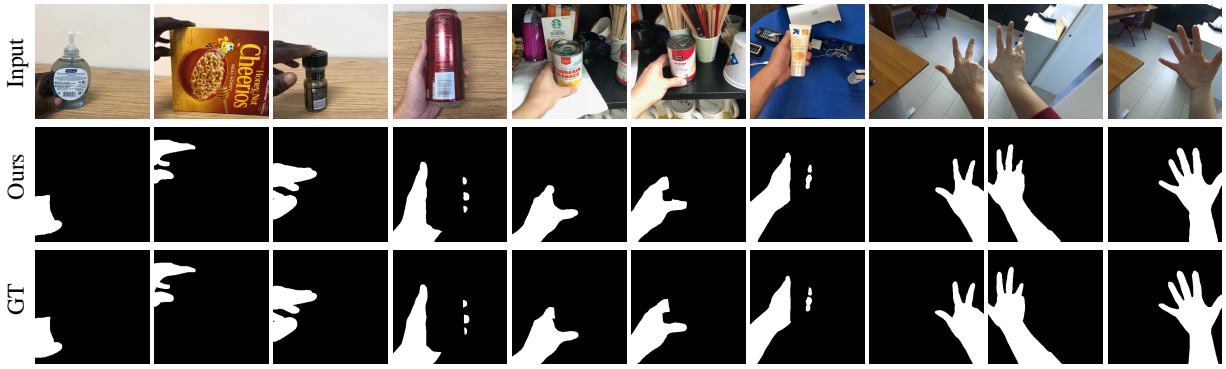

Figure 7: *Some qualitative results on test images (first row) from the upper limb segmentation dataset. Different scenarios, lighting conditions, skin tone, and hand-object occlusions are shown. The obtained predictions and the GT segmentation masks are displayed in the last two rows.*

DeepLabv3+ model. In particular, as reported in Section 3, the encoder component extracts low-level features and semantic information from the input RGB upper limb image, and the decoder computes the segmentation mask retrieving spatial and object boundary information. We evaluated the effectiveness of our deep learning method through both a quantitative and qualitative analysis, as explained in Section 6.

7.1. Limitations and Future Works

Using a simple 2D sprite in a 3D VE can cause the perception of the lack of depth of the limbs in the scene for the user. However, such a flatting effect can be mitigated by placing such a sprite in a corrected position in the user camera FOV. Another limit concerns the chromatic differences due to the different lighting conditions of real segmented upper limbs and the virtual scene. A possible solution may involve the use of image-to-image translation [IZZE16; CBC*19] and light style transfer [LCY*20; HMG20] techniques to adapt the real limb to the virtual scene. Moreover, segmentation mask errors can impact the visualization of users' upper limb in the VE. They could be mitigated by keeping the history of past frames. However, we assumed that these errors had little impact on the user experience. To this concern, we are preparing a controlled experiment to evaluate in the future the Usability, the User Experience, and the Sentiment of the participants through the well-known System Usability Scale (SUS) [Bro*96] and Self-Assessment Manikin (SAM) questionnaires [BL94]. In addition, we would like to repeat the comparative evaluation of freehand-steering locomotion techniques proposed by Caggianese et al. [CCE*20] by including our AV pipeline to assess the effectiveness of the growth of the user immersion feeling.

In the future, it would be also interesting to analyze the system by evaluating the sense of agency and body ownership and comparing it with other techniques. A possible issue may be due to the inconsistent display of the segmented hand in case the hand tracking from the Leap Motion fails as they are two independent phases. In this case, it could alter the sense of agency. Therefore, further investigation is needed or a different method for hand tracking should be provided. In particular, deep learning could also be used for hand tracking by retrieving hands joints positions using directly the images captured from simple RGB camera [MBS*18; GCEA20] and removing external infrared-based tracking devices such as Leap Motion. However, recent approaches [ZBV*20; CHS*19] raised several problems for the correct recognition of the depth component of the tracked hands. Indeed, they approximated the third dimension estimating the depth mask or evaluating the depth of the finger joints using the position of the wrist joint, defining the 2.5D. In future work, we would like to use our upper limb segmentation network to increase the RGB-only hand tracking approaches based on deep learning. We are confident that joint depth estimation can be improved by removing background from upper limb images. In this way, we can remove the Leap Motion controller from our proposed pipeline and then add a further deep learning step following our current segmentation step. Finally, our CNN method can also be used as a pre-processing step for hand gesture recognition [COS18] and left/right limb identification useful, for example, to associate specific actions to each hand.

Acknowledgment

The authors would like to thank NVIDIA's Academic Research Team for providing the Titan Xp cards under the Hardware Donation Program.

References

[19] *Deeplab Project*. 2019. URL: https://github.com/tensorflow/models/tree/master/research/deeplab (visited on 09/14/2021) 4.

[21] *Upper Limb Segmentation Web Page*. 2021. URL: http://graphics.unibas.it/www/EgoUpperLimbSeg_STAG/index.md.html (visited on 10/15/2021) 3, 4.

[AHTL16] ARGELAGUET, FERRAN, HOYET, LUDOVIC, TRICO, MICHAËL, and LÉCUYER, ANATOLE. *The role of interaction in virtual embodiment: Effects of the virtual hand representation. In 2016 IEEE Virtual Reality (VR). 3–10*. 2016 2.

[AWAN17] ABRAMOVICI, MICHAEL, WOLF, MARIO, ADWERNAT, STEFAN, and NEGES, MATTHIAS. "Context-aware maintenance support for augmented reality assistance and synchronous multi-user collaboration". *Procedia CIRP* 59 (2017), 18–22 2.

[BHB*20] BHANDARI, SIDDHARTH, HALLOWELL, MATTHEW R, BOVEN, LEAF VAN, et al. "Using augmented virtuality to examine how emotions influence construction-hazard identification, risk assessment, and safety decisions". *Journal of construction engineering and management* 146.2 (2020), 04019102 3.

[BL94] BRADLEY, MARGARET M and LANG, PETER J. "Measuring emotion: the self-assessment manikin and the semantic differential". *Journal of behavior therapy and experimental psychiatry* 25.1 (1994), 49–59 7.

[BLAL19] BORREGO, ADRIÁN, LATORRE, JORGE, ALCAÑIZ, MARIANO, and LLORENS, ROBERTO. "Embodiment and presence in virtual reality after stroke. A comparative study with healthy subjects". *Frontiers in neurology* 10 (2019), 1061 2.

[BOJ*09] BOHIL, CJ, OWEN, CHARLES B, JEONG, EJ, et al. "Virtual reality and presence". *21st century communication: A reference handbook* (2009), 534–544 2.

[Bro*96] BROOKE, JOHN et al. "SUS-A quick and dirty usability scale". *Usability evaluation in industry* 189.194 (1996), 4–7 7.

[BSRH09] BRUDER, GERD, STEINICKE, FRANK, ROTHAUS, KAI, and HINRICHS, KLAUS. "Enhancing presence in head-mounted display environments by visual body feedback using head-mounted cameras". *2009 International Conference on CyberWorlds*. IEEE. 2009, 43–50 2, 3.

[CBC*19] CAPECE, NICOLA, BANTERLE, FRANCESCO, CIGNONI, PAOLO, et al. "Deepflash: Turning a flash selfie into a studio portrait". *Signal Processing: Image Communication* 77 (2019), 28–39 7.

[CCE*20] CAGGIANESE, GIUSEPPE, CAPECE, NICOLA, ERRA, UGO, et al. "Freehand-Steering Locomotion Techniques for Immersive Virtual Environments: A Comparative Evaluation". *International Journal of Human–Computer Interaction* 36.18 (2020), 1734–1755 2, 4, 6, 7.

[CEGA20] CAPECE, NICOLA, ERRA, UGO, GRUOSSO, MONICA, and ANASTASIO, MARCO. "Archaeo Puzzle: An Educational Game Using Natural User Interface for Historical Artifacts". *Eurographics Workshop on Graphics and Cultural Heritage*. Ed. by SPAGNUOLO, MICHELA and MELERO, FRANCISCO JAVIER. The Eurographics Association, 2020. ISBN: 978-3-03868-110-6 2, 4.

[Cho17] CHOLLET, FRANÇOIS. "Xception: Deep learning with depthwise separable convolutions". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 1251–1258 3.

[CHS*19] CAO, Z., HIDALGO MARTINEZ, G., SIMON, T., et al. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) 7.

[COS18] CHALASANI, TEJO, ONDREJ, JAN, and SMOLIC, ALJOSA. "Egocentric Gesture Recognition for Head-Mounted AR Devices". *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 2018, 109–114 7.

[CPK*17] CHEN, LIANG-CHIEH, PAPANDREOU, GEORGE, KOKKINOS, IASONAS, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), 834–848 3.

[CPSA17] CHEN, LIANG-CHIEH, PAPANDREOU, GEORGE, SCHROFF, FLORIAN, and ADAM, HARTWIG. "Rethinking atrous convolution for semantic image segmentation". *arXiv preprint arXiv:1706.05587* (2017) 4.

[CZP*18] CHEN, LIANG-CHIEH, ZHU, YUKUN, PAPANDREOU, GEORGE, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". *Proceedings of the European conference on computer vision (ECCV)*. 2018, 801–818 3.

[DDF*18] DAMEN, DIMA, DOUGHTY, HAZEL, FARINELLA, GIOVANNI MARIA, et al. "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset". *European Conference on Computer Vision (ECCV)*. 2018 2.

[GCE21] GRUOSSO, MONICA, CAPECE, NICOLA, and ERRA, UGO. "Human segmentation in surveillance video with deep learning". *Multimedia Tools and Applications* 80.1 (2021), 1175–1199 6.

[GCEA20] GRUOSSO, MONICA, CAPECE, NICOLA, ERRA, UGO, and ANGIOLILLO, FRANCESCO. "A Preliminary Investigation into a Deep Learning Implementation for Hand Tracking on Mobile Devices". *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE. 2020, 380–385 7.

[GFG15] GÜNTHER, TOBIAS, FRANKE, INGMAR S, and GROH, RAINER. "Aughanded virtuality-the hands in the virtual environment". *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE. 2015, 157–158 2.

[GPK*18] GONZALEZ-SOSA, ESTER, PEREZ, PABLO, KACHACH, REDOUANE, et al. "Towards Self-Perception in Augmented Virtuality: Hand Segmentation with Fully Convolutional Networks." *Eurographics (Posters)*. 2018, 9–10 2, 3.

[GPT*20] GONZALEZ-SOSA, ESTER, PEREZ, PABLO, TOLOSANA, RUBEN, et al. "Enhanced Self-Perception in Mixed Reality: Egocentric Arm Segmentation and Database With Automatic Labeling". *IEEE Access* 8 (2020), 146887–146900 2, 3.

[GŞ18] GHEORGHIU, DRAGOŞ and ŞTEFAN, LIVIA. "Augmented Virtuality as an Instrument for a Better Learning of History". *13th International Conference on Virtual Learning (ICVL)*. 2018, 299–305 2.

[Hee92] HEETER, CARRIE. "Being there: The subjective experience of presence". *Presence: Teleoperators & Virtual Environments* 1.2 (1992), 262–271 2.

[HMG20] HART, DAVID, MORSE, BRYAN, and GREENLAND, JESSICA. "Style transfer for light field photography". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, 99–108 7.

[IZZE16] ISOLA, PHILLIP, ZHU, JUN-YAN, ZHOU, TINGHUI, and EFROS, ALEXEI A. "Image-to-Image Translation with Conditional Adversarial Networks. arxiv (2016)". *arXiv preprint arXiv:1611.07004* (2016) 7.

[KGS12] KILTENI, KONSTANTINA, GROTEN, RAPHAELA, and SLATER, MEL. "The sense of embodiment in virtual reality". *Presence: Teleoperators and Virtual Environments* 21.4 (2012), 373–387 2.

[LCBL16] LEE, GUN A, CHEN, JOSHUA, BILLINGHURST, MARK, and LINDEMAN, ROBERT. "Enhancing immersive cinematic experience with augmented virtuality". *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE. 2016, 115–116 2.

[LCY*20] LIU, TONG, CHEN, ZHAOWEI, YANG, YI, et al. "Lane detection in low-light conditions using an efficient data enhancement: Light conditions style transfer". *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2020, 1394–1399 7.

[LK13] LI, CHENG and KITANI, KRIS M. "Pixel-level hand detection in ego-centric videos". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, 3570–3577 3.

[LK19a] LEE, KYUNGJUN and KACORRI, HERNISA. "Hands Holding Clues for Object Recognition in Teachable Machines". *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019 3.

[LK19b] LI, GEN and KIM, JOONGKYU. "DABNet: Depth-wise Asymmetric Bottleneck for Real-time Semantic Segmentation". *British Machine Vision Conference*. 2019 3.

[LMB*14] LIN, TSUNG-YI, MAIRE, MICHAEL, BELONGIE, SERGE, et al. "Microsoft coco: Common objects in context". *European conference on computer vision*. Springer. 2014, 740–755 4.

[LR19] LATEEF, FAHAD and RUICHEK, YASSINE. "Survey on semantic segmentation using deep learning techniques". *Neurocomputing* 338 (2019), 321–348 6.

[LSD15] LONG, JONATHAN, SHELHAMER, EVAN, and DARRELL, TREVOR. "Fully convolutional networks for semantic segmentation". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, 3431–3440 3.

[MBMB15] MCGILL, MARK, BOLAND, DANIEL, MURRAY-SMITH, RODERICK, and BREWSTER, STEPHEN. "A dose of reality: Overcoming usability challenges in vr head-mounted displays". *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015, 2143–2152 3.

[MBP*21] MINAEE, SHERVIN, BOYKOV, YURI Y, PORIKLI, FATIH, et al. "Image segmentation using deep learning: A survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) 6.

[MBS*11] MCMANUS, ERIN A, BODENHEIMER, BOBBY, STREUBER, STEPHAN, et al. "The influence of avatar (self and character) animations on distance estimation, object interaction and locomotion in immersive virtual environments". *Proceedings of the ACM SIGGRAPH Symposium on applied perception in graphics and visualization*. 2011, 37–44 2.

[MBS*18] MUELLER, FRANZISKA, BERNARD, FLORIAN, SOTNYCHENKO, OLEKSANDR, et al. "GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB". *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. June 2018 7.

[Min95] MINE, MARK R. *Virtual Environment Interaction Techniques*. Tech. rep. USA, 1995 4.

[MTRW17] MCFADDEN, D, TAVAKKOLI, A, REGENBRECHT, J, and WILSON, B. "Augmented Virtuality: A Real-time Process for Presenting Real-world Visual Sensory Information in an Immersive Virtual Environment for Planetary Exploration". *AGU Fall Meeting Abstracts*. Vol. 2017. 2017, IN32B–07 2.

[MTUK95] MILGRAM, PAUL, TAKEMURA, HARUO, UTSUMI, AKIRA, and KISHINO, FUMIO. "Augmented reality: A class of displays on the reality-virtuality continuum". *Telemanipulator and telepresence technologies*. Vol. 2351. International Society for Optics and Photonics. 1995, 282–292 2.

[NAA18] NEGES, MATTHIAS, ADWERNAT, STEFAN, and ABRAMOVICI, MICHAEL. "Augmented Virtuality for maintenance training simulation under various stress conditions". *Procedia Manufacturing* 19 (2018). Proceedings of the 6th International Conference in Through-life Engineering Services, University of Bremen, 7th and 8th November 2017, 171–178. ISSN: 2351-9789 2.

[PD19] PIGNY, PIERRE-OLIVIER and DOMINJON, LIONEL. "Using cnns for users segmentation in video see-through augmented virtuality". *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE. 2019, 229–2295 2, 3.

[PKS15] PAPANDREOU, GEORGE, KOKKINOS, IASONAS, and SAVALLE, PIERRE-ANDRÉ. "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, 390–399 3.

[PPC21] PALMA, GIANPAOLO, PERRY, SARA, and CIGNONI, PAOLO. "Augmented Virtuality Using Touch-Sensitive 3D-Printed Objects". *Remote Sensing* 13.11 (2021), 2186 2.

[RAS19] RAUTER, MICHAEL, ABSEHER, CHRISTOPH, and SAFAR, MARKUS. "Augmenting virtual reality with near real world objects". *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2019, 1134–1135 2.

[RDS*15] RUSSAKOVSKY, OLGA, DENG, JIA, SU, HAO, et al. "Imagenet large scale visual recognition challenge". *International journal of computer vision* 115.3 (2015), 211–252 4.

[RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. "U-net: Convolutional networks for biomedical image segmentation". *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, 234–241 3.

[RFLF21] RAGUSA, FRANCESCO, FURNARI, ANTONINO, LIVATINO, SALVATORE, and FARINELLA, GIOVANNI MARIA. "The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, 1569–1578 2.

[RLK*04] REGENBRECHT, HOLGER, LUM, TIM, KOHLER, PETRA, et al. "Using augmented virtuality for remote collaboration". *Presence* 13.3 (2004), 338–354 2.

[ROW*03] REGENBRECHT, HOLGER, OTT, CLAUDIA, WAGNER, MICHAEL, et al. "An augmented virtuality approach to 3D videoconferencing". *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* IEEE. 2003, 290–291 2.

[SLD*18] SCHWIND, VALENTIN, LIN, LORRAINE, DI LUCA, MASSIMILIANO, et al. "Touch with foreign hands: The effect of virtual hand appearance on visual-haptic integration". *Proceedings of the 15th ACM Symposium on Applied Perception*. 2018, 1–8 4.

[ZBV*20] ZHANG, FAN, BAZAREVSKY, VALENTIN, VAKUNOV, ANDREY, et al. "Mediapipe hands: On-device real-time hand tracking". *arXiv preprint arXiv:2006.10214* (2020) 7.