AR-MoCap: Using augmented reality to support motion capture acting

AR-MoCap: Using Augmented Reality to Support Motion Capture Acting

Alberto Cannavò* Filippo Gabriele Prattichò† Alberto Bruno‡ Fabrizio Lamberti§

Politecnico di Torino, Dipartimento di Automatica e Informatica, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy

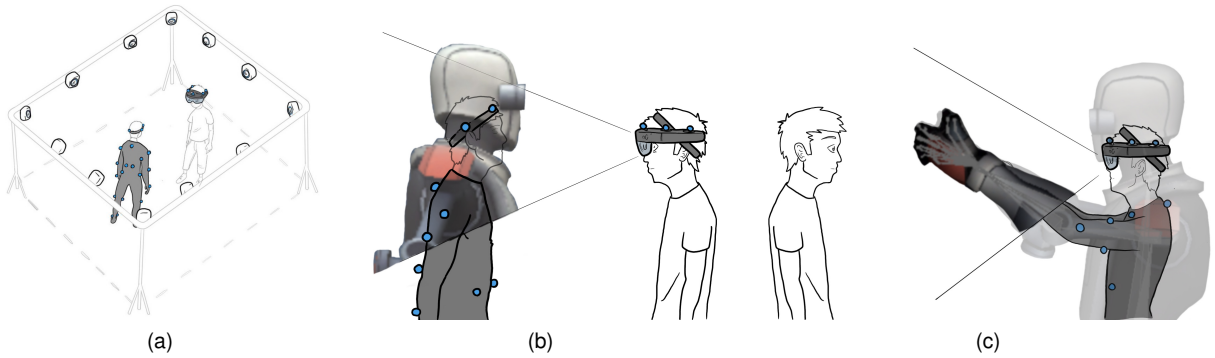


Figure 1: Conceptualization of the AR-MoCap system functioning and supported use cases: a) rehearsal and shooting environment, b) augmentation of other characters, and c) augmentation of the own character.

ABSTRACT

Technology is disrupting the way films involving visual effects are produced. Chroma-key, LED walls, motion capture (mocap), 3D visual storyboards, and simulcams are only a few examples of the many changes introduced in the cinema industry over the last years. Although these technologies are getting commonplace, they are presenting new, unexplored challenges to the actors. In particular, when mocap is used to record the actors' movements with the aim of animating digital character models, an increase in the workload can be easily expected for people on stage. In fact, actors have to largely rely on their imagination to understand what the digitally created characters will be actually seeing and feeling. This paper focuses on this specific domain, and aims to demonstrate how Augmented Reality (AR) can be helpful for actors when shooting mocap scenes. To this purpose, we devised a system named AR-MoCap that can be used by actors for rehearsing the scene in AR on the real set before actually shooting it. Through an Optical See-Through Head-Mounted Display (OST-HMD), an actor can see, e.g., the digital characters of other actors wearing mocap suits overlapped in real-time to their bodies. Experimental results showed that, compared to the traditional approach based on physical props and other cues, the devised system can help the actors to position themselves and direct their gaze while shooting the scene, while also improving spatial and social presence, as well as perceived effectiveness.

Keywords: Collaborative virtual production, acting rehearsal and performance, motion capture, body ownership, virtual characters, visual effects, augmented reality

Index Terms: Human-centered computing—Human computer interaction (HCI)—Mixed / augmented reality—; Human-centered computing—Human computer interaction (HCI)—Collaborative interaction—; Computing methodologies—Computer vision—Image and video acquisition—Motion capture; Human-centered computing—Applied computing—Arts and humanities—Performing arts; Human-centered computing—Applied computing—Arts and humanities—Media arts;

*e-mail: alberto.cannavo@polito.it
†e-mail:filippogabriele.prattico@polito.it
‡e-mail: alberto.bruno@studenti.polito.it
§e-mail: fabrizio.lamberti@polito.it

1 INTRODUCTION

Technological advancements are changing the way how movies are produced [21]. Indeed, the most recent technical turns regard computer generated imagery (CGI) and visual effects (VFX) [2]. In fact, the use of CGI and VFX has become prominent not only for the production of science-fiction movies, but also of many other genres [13]. Despite their widespread adoption, however, such technologies are also presenting new challenges to both technical and acting crews [12].

1.1 Virtual Production Pipeline

The production of movies containing VFX can represent a very laborious and complex task, involving the collaboration of different roles and departments like, e.g., director, VFX supervisor, animators, actors, production and art departments, etc.

According to the guidelines proposed by the famous, award-winning VFX supervisor Andrew Whitehurst, the production pipeline can be regarded as split in three phases: pre-production, production, and post-production [29]. The pre-production phase is characterized by all the activities carried out before the shooting actually starts on set. This stage includes the following steps: i) research and development, i.e., the implementation or optimization of tools aimed at satisfying the needs related to the movie production; ii) testing, i.e., the creation of proofs of concept for demonstrating potential look, style, or possible technologies to be leveraged in the production; iii) previsualization (previs), i.e., the development of 3D representations of the script/storyboard using low-resolution models and textures; iv) early production of assets, where digital and 3D assets for the backgrounds and/or the characters start to be created, thus becoming available for the next steps.

In the production phase, the actual shooting of the scenes is performed. Members of the VFX team are present on the set to take photographs that will be used as references for modeling, texturing, and lighting. 3D scanning techniques are used to reconstruct props, environments, and buildings. If motion capture (mocap), i.e., the

process of recording a live motion event by tracking key points in space over time on an actor's body or face, a prop, etc. [18] is considered, members of the team are present on the set to arrange the required equipment (including mocap suits, tracking cameras, head-mounted devices for facial capture, etc.).

The post-production phase includes several steps aimed to finalize the movie production, like adding 3D animations (e.g., of modeled and rigged objects) and physics-based effects (simulating particle systems, rigid-body dynamics, fluids, etc.), configuring and running the rendering (setting up the lighting and producing the finished CGI), performing the compositing (combining all the elements to create a seamless finished image that looks as if it had been filmed and no syntehtic effects have been added), etc. 3D scans and camera tracking information gathered in the production phase are used in post-production to add the CGI elements.

In recent years, however, the introduction of new technologies in the production process is altering more and more the original structure of the pipeline [20]. These changes are leading to a so-called "virtual production". As stated by Kadner et al. ([11]), this term refers to a spectrum of computer-aided production and visualization film-making methods that, by combining Augmented and Virtual Reality (AR/VR) with CGI and game engine technologies, allows producers to visualize their scenes unfold directly while they are assembled and captured on-set. Besides introducing a paradigm shift that is transforming the linear process of the traditional pipeline, with tasks well separated among the departments, into a more interactive process characterized by joint work, virtual production is also increasing the efforts made during the preparation activities.

Thus, previs, which is now supported by innovative technologies such as the mentioned AR/VR, simulcams, and LED walls, is becoming of paramount importance in movie production, as it makes people involved in the overall process more aware of the final result. Improvements in the field of previs let directors and other key members of the staff understand how the film is going on, by sharing a common vision and limiting possible misunderstandings. By simply wearing an headset, directors and actors can see 3D reconstructed environments that are representative of what it might actually get filmed in the final product. Thus, the actors' performance can be enhanced, and directors can make decisions by analyzing more reliable outputs. Assets created for the previs (e.g., 3D models, rigged characters, animations, etc.) can also be used later in the production without the need to recreate them since, nowadays, it is possible to handle high-quality resources and dynamically degrade them when necessary. Moreover, high-quality previs can also be leveraged to study the reaction of the audience, taking notes of what worked and what should be improved in terms of storyboard, shots, lighting, etc.

1.2 Challenges for Actors

The benefits provided by a better visualization of the scenes thanks to the use of new technologies became more evident in the production of movies containing mocap shoots. In this way of acting, recorded movements are leveraged to animate computer-generated contents. These contents are used to "augment" the real environment with synthetic elements (i.e., 3D avatars and/or objects) that are added to the scene after shooting. In this way, actors are requested to perform by imagining virtual environments that could be significantly different from what they are actually seeing at shooting time, as computer-generated assets are added only at a later time, i.e., in post-production phase.

Actors' performance can get even worse when they are asked to interact with other actors portraying characters who do not match their appearance or are totally virtual [12]. In fact, with mocap it is possible to bring to life characters that are non-antropomorphic, and/or present body proportions and sizes totally different from the actors who are portraying them. This leads to an additional complexity for the actors, who have to empathize with fictional characters, thinking of their real body as the character's body and animating the virtual character only seeing their own, real movements. Many actors lament the fact that this way of acting is very far from what they have learned in drama schools, and makes them frustrated [26]. Difficulties in shooting these kinds of scenes get particularly evident in the production phase, when the actual scene context is added, as any miscommunication or misunderstanding leads to inconsistencies between the performance and the context [12].

The separation between what the actors of mocap scenes are performing on set and what the scene actually contains after production and post-production causes many issues that they have to deal with while shooting or rehearsing the scenes. Issues regard, e.g., the movements of the actors that are not aligned with the desired appearance or are not properly reacting to the surroundings. As said, acting gets even more demanding when the physiognomies of the actors are largely different from their digital representations. Typically, in these scenes, actors are helped by mechanisms designed to direct their gaze and performance towards placeholder props, indicating where and how the digital counterpart is actually located and shaped. Unfortunately, depending on the character, the use of these mechanisms may not be feasible or effective due to physical constraints [12]. Moreover, physical props do not allow mocap actors to direct/keep the eye contact with the other actors during the performance, as the latter may have to be kept in their peripheral vision to properly control the character. In scenes requesting the actors to maintain a consistent eye-line towards moving virtual objects/characters, laser pointing is also exploited. However, respecting the exact timing of the movement to be followed could be difficult to achieve as, generally, animations are pre-computed [2]. This may result in imprecise acting that can have unfortunate consequences not only in the shooting phase (as it increases the time needed to shoot the scene), but also in post-production. In fact, mismatches of the actors' gaze typically lead to extensive post-processing workload aimed to recreate or modify the animations to match as much as possible what has been filmed [2].

1.3 Opportunities for AR/VR

AR and VR could, in principle, help to address some of the above issues. Indeed, their use in the cinema industry has already been explored [7]. For instance, they have been used to support filmmaking [15, 19] and digital storytelling [23], the configuration/visualization of sets [24], the pre-visualization of scenes [10, 28], etc.

In the context of mocap acting, AR/VR technologies can be exploited to let the actors see a scene that is more similar to what it will look like in the final product. In this way, the actors' contextual awareness could grow, as they would be allowed to visualize digital contents without the need to imagine them, and this may improve their performance [27]. The use of AR/VR can not only help the actors to act more intuitively and adequately while shooting or rehearsing the scene [2], but also lead to lower post-production efforts, since it gives more believability to the actors' performance [27]. For instance, the possibility to perceive the actual size of the virtual characters makes it easier for the actors to keep eye contact and interact with them [2]. These technologies also help the actors to empathize better with the surroundings and create emotional connections with the characters since, if they can see what actually is in the scene, they can also interpret emotions in a more natural and real way (e.g., be scared of another character) [7]. The improved awareness within the scene derived from the introduction of AR/VR can also be leveraged by the directors to better express their creative intent and plan sequences shooting on-set [2].

Considering the above aspects, it is not surprising to find in the literature works proposing VR-/AR-based systems for mocap scene rehearsal. The majority of these works envisaged VR-based solutions, as their goal was to work with fully digital environments. To the best of the authors' knowledge, AR-based solutions supporting

mocap acting have not been investigated yet. The use of AR, however, could bring a number of advantages over VR. For instance, with AR it would be possible to remove the need to model and track real objects if they have to be used as props in the scene. Moreover, some actions would be easier to simulate (e.g., knotting a rope receiving the correct haptic feedback) or perform (e.g., climbing a stair) in AR than in VR.

1.4 Contribution

Based on the considerations expressed above, this paper aims to present and investigate the effectiveness of AR-MoCap (Fig. 1), an AR-based system designed to support actors when rehearsing or shooting mocap scenes involving VFX.

By wearing an Optical See-Through Head-Mounted Display (OST-HMD), i.e, an AR headset that makes it possible to superimpose digital contents on the real-world view while optically maintaining a see-through view, the actors are allowed to visualize in real time 3D virtual avatars. The virtual representations of the avatars are superimposed on the actors who are controlling them through mocap suits. Besides visualizing the avatars, the actors are allowed to see the other real and virtual elements of the scene.

The assumptions behind the design of AR-MoCap are that, by using this system, the actors can gain a better familiarity with the scene being acted as a consequence of an improved sense of spatial and social presence. In this respect, it is also considered the positioning in the environment and interaction with the virtual characters, both those intended to be added and animated just in post-production and those controlled by actors via mocap. This increased context awareness is expected to lead to less animation cleanups (e.g., for fixing gaze mismatches) and to a higher quality shooting, overall.

To confirm the assumptions, a user study was carried out by collecting both objective and subjective measurements. Experimental results showed that the devised approach helps the actors to direct their gaze and position on stage better than with traditional techniques for scene rehearsal based on physical props and visual cues. Moreover, the approach improves the spatial and social presence, as well as the perceived effectiveness of the rehearsal method.

2 Related Work

In the following, works concerning tackled technologies as well as addressing the challenges of virtual production are reviewed.

2.1 Mocap and AR/VR Technologies

The possibility to combine mocap and AR/VR technologies is a well-explored field in the literature. An example in the context of training is provided, e.g., in [4]. This work presents a VR system that can be used as a self-learning tool to improve the execution of a basketball-related technical gesture. A mocap suit is exploited to reconstruct in VR the real-time skeleton data representing the player's arm, helping him or her to replicate the reference gesture shown using the ghost metaphor, i.e., a VR-based motion-guiding interface used to visualize a 3D reconstruction of the reference gesture in the virtual environment. Similarly, the works in [5] and [16] present training systems for improving Tai Chi movements. Mocap suits are used to record and track in real time the movements of both a trainer and a trainee. Within a VR environment, the trainee can receive feedback regarding the execution of his or her movements and visualize the correct ones, which need to match those of the trainer avatar. For what it concerns the use of AR, the work in [9] proposes a self-learning tool for improving golf movements. The trainee can visualize pre-recorded movements performed by a trainer avatar by wearing a Microsoft's HoloLens device. Similarly, the work in [6] combines an OST-HMD and a mocap suit to show a reconstructed avatar, enabling real-time, full-body interactions in AR.

2.2 Mocap and AR in the Cinema Industry

The works mentioned above represent good examples of how to combine the considered technologies. Their focus, however, is mainly on (sport) training. An example of works using AR/VR technologies in the cinema industry is represented by [24]. The authors present an AR system designed to let directors validate the final setup of the set. To this aim, directors can visualize and manipulate computer-generated assets placed in the real environment by making use of an OST-HMD (for seeing the assets) and a tablet (for manipulating them). The proposed approach lets directors save time and effort in testing different configurations, since they can be seamlessly recreated as virtual scenes before actually placing the real objects in the environment. Another example showing the application of these technologies in the cinema industry is reported in [27]. The work presents an Android app for the production of low-cost movies that allows the actors to switch between visualizing the real environment (that contains green screen areas) and the synthetic environment that is generated by superimposing digital contents over the green areas. The work in [10] describes a previs method that uses AR to let videographers test the movements of the camera used to shoot scenes involving real actors without the need for their presence. Virtual avatars are superimposed on live videos in real-time considering the actual position and orientation of the camera. The core of the devised method is an algorithm for estimating the motion of the camera in real-time based on image data. Similarly, the work in [28] describes how computer vision can support the previs of scenes that take place in open sets or at outdoor locations, and in which a real background is combined with computer-generated contents (human characters and other creatures). A vision-based camera tracking method is proposed that leverages environmental information to improve the estimation of camera position and orientation, thus achieving better results in the superimposition of digital contents; 3D video obtained from multiple cameras is also used to capture the actors' movements.

2.3 Scene Rehearsal in AR/VR

Moving to solutions supporting scene rehearsal, it is noticeable the increasing interest in technologies such as LED walls, i.e., immersive and massive video walls in which physical set objects are combined with digital extensions on screens. This technology (that has been exploited already in productions like, e.g., "The Mandalorian" (2019)[1]) can be used not only for practicing but also for live filming, removing the need for location shoots [25]. Despite the promising benefits, LED walls are still characterized by additional efforts required, e.g., for ensuring a perfect match between the real and virtual sets, correcting volumetric 3D colors and lighting, staging of set objects to hide the edges of the LED screens, etc. [25].

Alternative solutions based on AR/VR were proposed in the literature to support actors in practicing the scenes. For instance, the work in [2] presents a system helping actors to rehearse VFX-enhanced scenes. In this case, VR is leveraged not only to make actors feel immersed in the digital scenarios, but also to provide them with "dynamic scenario" features; these features are designed to allow them interact with virtual elements (e.g., take a glass, move a chair, switch a light on, etc.) while rehearsing dialogue and action at their own speed. The system supports both single- or multi-user operations, and it can be used both for on-set and off-set rehearsals. The work in [12], in turn, addresses the difficulties faced by the actors when shooting motion-capture scenes that involve virtual characters of different scale sizes. To this aim, the authors developed an immersive VR system that lets the actors visualize the body of their virtual avatar while seeing also the other characters (having different sizes) from their own viewpoint.

Although VR supports high-quality simulations and graphics, its use for scene rehearsal may introduce some drawbacks. First, in

[1]The Mandalorian: `https://bit.ly/3GPuBL5`

Table 1: Overview of related work.

	AR/VR tech.	Mocap	Field
[4]	VR	✓	Sport training
[5, 16]	VR	✓	Sport training
[9]	AR (OST-HMD)	✓	Sport training
[6]	AR (OST-HMD)	✓	Training & rehabilit.
[24]	AR (OST-HMD)		Previs
[27]	AR (Mobile dev.)		Previs
[10, 28]	MR		Previs
[2]	VR (CAVE)		Rehearsal
[12]	VR (HMD)	✓	Rehearsal
[8]	AR (OST-HMD)		Rehearsal
AR-MoCap	AR (OST-HMD)	✓	Rehearsal

the case of scenes requiring interactions with elements of the real environment, an intense modeling/reconstruction effort is necessary in order to prepare the required assets; sophisticated techniques also need to be used to track them. Moreover, some interactions (like, e.g., touching a fluid or knotting a rope) could be difficult to reproduce in VR; the same would apply to actions requiring the actors to take advantage of the scenography (like, e.g., climbing a stair). Another possible drawback regards the possibility to keep eye contact with the other actors (not embodying virtual characters) or crew members while acting without the need to unwear the headset, as eye contact is regarded as essential to prime emotional bond or receiving feedback during the performance [12].

The above issues could be mitigated by leveraging AR, rather than VR. An example of AR usage in the considered domain is provided by [8]. The work describes a Mixed Reality (MR) scene rehearsal system in which virtual characters can be visualized through an OST-HMD from a first-person view. In this way, the actors can practice the scene before actually shooting it, possibly interacting with virtual characters. The scene considered as a use case represents a battle between two samurai: one real (the actor wearing the HMD), and one virtual. By tracking the sword of the real actor, it is possible to let him or her interact with the enemy and also provide vibrotactile feedback when the swords clash.

2.4 Summary

Table 1 summarizes the relevant works reviewed above, reporting used technologies and fields of application. Considering the aspects tackled so far, as well as the advantages and drawbacks of proposed approaches, the AR-MoCap system presented in this paper leverages mocap and AR (through an OST-HMD) to let actors rehearse scenes including VFX. From a technological viewpoint, the proposed system is related to [6,9], as it combines the visualization of mocap data through an AR OST-HMD, but for a different purpose. Moreover, the devised approach shares similarities in terms of application scenarios with previous works in the cinema industry. More specifically, like [2, 12], AR-MoCap supports actors in the rehearsal of scenes including VFX: differently than these works, though, the proposed system relies on AR to convey digital contents, since a number of issues still remain when leveraging VR.

3 Materials and Methods

This section first illustrates the possible ways in which AR can be exploited to help actors preparing to shoot mocap scenes that involve post-produced characters. Afterwards, it presents the AR-MoCap system, which leverages AR to support rehearsal (Fig. 1a). Finally, it illustrates the user study that, following the methodology adopted in [2,12], has been devised to assess the effectiveness of the proposed approach against traditionally employed methods.

3.1 Use Cases

The AR-MoCap system is meant to be used for scenes in which there is at least one actor for each of the following roles: one actor that portrays real characters (no substantial post-production will be applied to rework the shooting of the real actor), and another actor that controls, with his or her movements, a virtual character by means of mocap. As said, the use of AR makes it possible to merge elements of the real environment with computer-generated assets before passing to the post-production phase; hence, scenes to be rehearsed/shooted may include elements whose size could change in the final product. Interactions with real, physical objects are considered too. Within this context, the spectrum of possible application scenarios for the AR-MoCap system can be identified by analyzing two extreme use cases:

A) *Augmenting the other characters in the scene* (Fig. 1b): During the rehearsal of the scene, the actor who is playing the real character wears the HMD. In this way, he or she is allowed to see the virtual characters (superimposed on the actors who are wearing the mocap suit) and the surrounding real environment.

B) *Augmenting the own character* (Fig. 1c): During the rehearsal, the actor who plays the virtual character by controlling it through mocap also wears the HMD. In this way, he or she can observe the virtual character's body parts superimposed on his or her own parts, as well as the additional parts that may belong only to the virtual character like, e.g., wings or arm extensions, etc. Differently than in the previous use case, in which post-production operations are needed to remove from the image the HMD worn by the actor (if he or she is playing a real character), in this case, the body of the actor has to be fully replaced with that of the virtual character which is controlled with mocap, thus enabling the possibility to use the proposed system also during the shooting. It is speculated that the actor's performance can improve as a result of increased body ownership and control over the controlled virtual character, both in terms of self-awareness and in relation to other actors and scene elements.

Even though both the use cases look promising, the second one would be more affected by the limited field-of-view of common OST-HMDs, which would make it difficult for the actor to see all the portions of his or her virtual character (especially at a close distance). Thus, in presenting the architecture and performing the experimental evaluation, the first use case and an OST-HMD were considered, even though the AR-MoCap system can support also the other use case and HMD technology. Indeed, the field-of-view limitation could be mitigated by using a video see-through HMD (VST-HMD) like the Varjo-XR3 or Meta Quest Pro. However, a VST-HMD would obstruct the face of the actor wearing it more than an OST-HMD, preventing the other actors in the scene to engage in eye-contact and see facial expressions [12] with the one wearing the headset (important for B) or viceversa (relevant for A). Even though eye- and facial-tracking capabilities are starting to be integrated into VST-HMDs, these technologies are still immature, and their impact on the emotional expressivity and engagment is yet to be verified in scenarios like the one tackled in the present work.

3.2 AR-MoCap Architecture

As anticipated, the goal was to devise a system capable to let an actor wearing an OST-HMD visualize in AR the digital contents that are supposed to be added in the post-production phase (or, at least, a simplified version of them, if not available yet).

These contents include digital characters animated either via pre-recording or in real-time using mocap. The high-level architecture of the devised system is depicted in Fig. 2. As OST-HMD, it was used

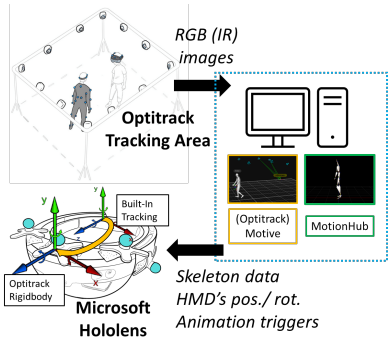


Figure 2: Architecture of the AR-MoCap system.

the HoloLens (1st Gen)[2]. In order to deliver the AR contents, an application was implemented using the Unity game engine (v2020.3 LTS). The application was populated with the necessary 3D models together with the pre-recorded animation and sound effects through spatialized audio) required to arrange the intended screenplay. It is worth obsering that using the AR-MoCap system entails a limited manual effort, i.e., the definition of the scene logic (easily accomplished with a game engine), as 3D assets generated for previs can be used also to this purpose.

The sequence of animations was implemented using the Unity timeline to allow the actor (or the experiment administrator in the role of a director, for the user study) to trigger them in the right order with a single input provided by the HoloLens clicker device.

The mocap was performed using a 8 RGB(IR)-camera (*Prime 13W*) Optitrack system[3] in a tracked space of 4.5m×6.0m. Mocap was exploited both to animate the character in real time (using a mocap suit worn by the actor) and to perform positional and rotational tracking of the HMD.

Actor's skeleton data and HMD tracking data were streamed from the PC hosting the Optitrack software (Motive v2.2) to the HoloLens application over a 2.4GHz Wi-Fi channel. Since the Optitrack streaming library (NatNet SDK) is not supported by the HoloLens, an additional middleware software was used to encapsulate the mocap data coming from the Motive software and stream them using the UDP protocol to the HoloLens application. Specifically, the open-source MotionHub middleware [14] was employed. The MotionHub software (hosted on the same machine of Motive) was customized by adding the support to stream rigidbodies (i.e., the HMD tracking data) and modifying the user interface by adding a button to trigger the next animation in the HoloLens application. This implementation was preferred to simply using the HoloLens clicker straight into the application, so to potentially enable multiple users (e.g., other actors or the director) to activate the next animation.

The motion-to-photon latency (the time interval between the movement of the real actor and the rendering of the virtual character copying it) of the system was measured to be around 50ms. In this respect, it is worth mentioning that the HoloLens tracking data provided by the Optitrack system were not employed to fully override the built-in tracking, since the additional latency could have emphasized the misalignment between the real world and the registered AR contents, especially in case of fast head movements; thus, they were only exploited to align the two reference systems and periodically correct the drift.

Finally, in order to support the experimental evaluation, the HoloLens application was endowed with the features required to collect the metrics reported in the following, as well as to support a functioning mode in which the digital contents are not rendered and the audio can be played back into an external Bluetooth speaker system available in the shooting room. In this way, it was possible to use the HMD just as a measuring tool, without providing visual cues for acting, but anyhow providing the actors with the audio cues at rehearsal/shooting time.

3.3 Scene and Script Design

As in the study reported by Kammerlander et al. [12], to run the user study a custom script was created for the scene rehearsal, which contains a number of possible difficulties that actors may face when shooting scenes in the considered use case. In particular, the difficulties that were chosen to be stressed are: directing the actors' gaze on specific virtual elements (e.g., to part of the virtual character's body or virtual objects), positioning in the environment, and reacting emotionally to events involving virtual objects/characters. Moving from these considerations, a scene (and a script) was designed by taking into account the following aspects: i) the scene should contain at least one animated virtual character controlled by mocap; ii) the size of that virtual character should be different than that of the actor playing such role; iii) the scene should require the actor to interact with virtual characters and objects; the objects can be either physical props belonging to the real environment or synthetic, computer-generated assets; iv) emotional reactions should be required from the actor with respect to events involving virtual objects/characters.

In order to create a scene (script) fulfilling all these requirements, several recent movies presenting the characteristics presented above (e.g., "The One and Only Ivan" (2020)[4], "War for the Planet of the Apes" (2017)[5], Marvel's films like "Avengers: End Game" (2019)[6], "The Hobbit" (2014) [7], etc.) were analyzed. Based on this analysis, a scene from the movie "Aladdin" (2019)[8] was chosen as a reference. This scene was considered as particularly suited to the purpose, since it demands continuous interactions between two actors, one of whom controls a virtual character by means of mocap. Moreover, the virtual character changes the body size several times. Finally, the scene includes interactions with real and virtual objects.

The scene involves two characters (a guy and a genie) and, for the experiments, the system was configured to support the considered use case (A): more specifically, the guy was played by the study participant wearing the OST-HMD, whereas the genie was portrayed by a single operator who wore the mocap equipment. An additional character, i.e., a hell dog, was added to the scene with the aim of stimulating strong emotional reactions in the actor due to its aggressive aspect and behavior, as suggested in [12]; the movements of this character were activated programmatically.

During the scene (whose salient moments are depicted in Fig. 3), the character played by the participant helps the genie to free himself from a cage. To accomplish this task, a number of actions (e.g., tearing off a page from a spell book, sketching on a paper, etc.), interactions (i.e., cutting a rope to which a hell dog is hooked, throwing a book page to the other character, selecting one of two fluctuating objects, following with the gaze the movements of the other character), and emotional reactions (i.e., being frightened of the hell dog and surprised by the transformations happening to the other character) are required.

In the scene, the genie assumes different sizes (starts at 2.10m tall, then downsizes to 50cm) with the aim to stress the actor's behavior of directing his or her gaze to/interact with a virtual character played by an actor of a different size, which is a core aspect under investigation. The numerous interactions with the virtual objects enable

[2]HoloLens (1st Gen): `https://learn.microsoft.com/en-us/hololens/hololens1-hardware`

[3]Optitrack: `https://optitrack.com/cameras/primex-13w/`

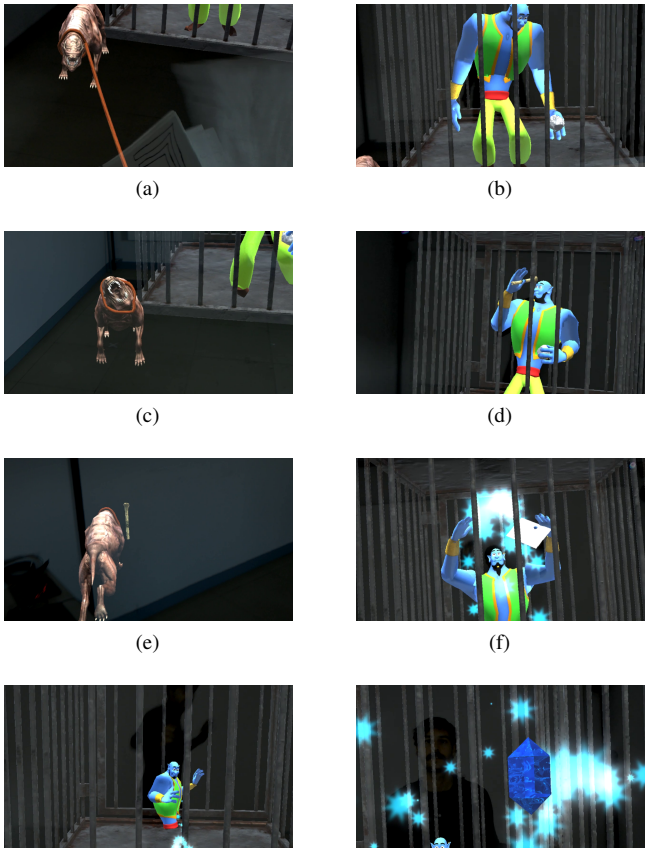
[4]The One and Only Ivan: `https://bit.ly/3HbkBgA`

[5]War for the Planet of the Apes: `https://bit.ly/3iEIKCZ`

[6]Avengers End Game: `https://bit.ly/3Wc03ZK`

[7]The Hobbit: `https://bit.ly/3IRHMxO`

[8]Aladdin: `https://bit.ly/3CW3Q6G`



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3: Salient moments of the devised script from the actor perspective (AR cues): a) a page from the magic book is teared off by the actor and crumpled with the dog still on the leash; b) the crumpled page is thrown to the genie; c) the hell dog rope is cut; d) the genie finds a bone and throws it far away; e) the dog chases the bone; f) the genie casts the spell to shrink himself; g) the genie jumps out of the cage; h) the genie spawns two reward diamonds, which fluctuates in the scene, and the blue diamond is grabbed by the actor.

the participant to experiment with positioning issues (for instance, to cut the rope a consistent position has to be assumed). It is worth observing that, in order to support the relevance of AR, the script includes actions that would be difficult to faithfully reproduce in VR (leafing a spell book, tearing off a page from it, throwing it, and writing notes).

The script is available at `http://tiny.cc/armocap_script`.

4 Experiment

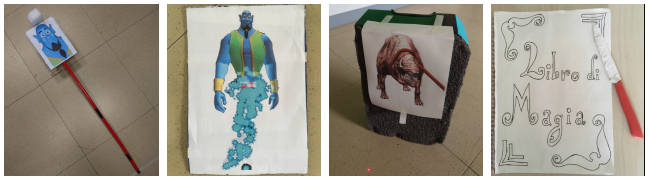
This section reports on the user study performed in order to evaluate the proposed AR-MoCap system for the use case A.

4.1 Study Design and Tools

The study design and the procedure adopted for the experimental evaluation was inspired by [2]. The aim was to compare two different methods of rehearsing scenes involving mocap: the traditional method based on physical props and laser pointing (later referred to as TR) and the proposed AR-based one (referred to as AR). Fig. 4 shows the use of the two methods for the considered scene. Videos are also available at `http://tiny.cc/armocap_videos`.

(a) (b)

Figure 4: Methods adopted in the experimental evaluation for scene rehearsal: a) TR, and b) AR.



(a) Tall genie (b) Tiny genie (c) Hell dog (d) Magic book and knife

Figure 5: Physical props used in the TR method.

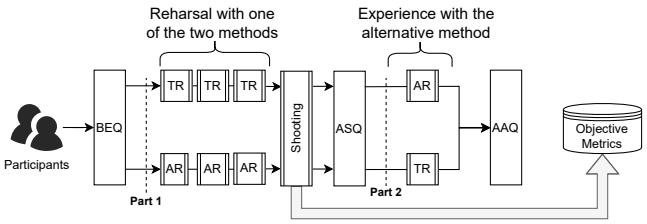


Figure 6: Study design.

TR rehearsal: the participants practiced the interactions with the virtual characters and the elements of the digital environment by means of physical props representing them (Fig. 4a). More specifically, cardboard figures depicting the genie were used to provide the participants with a reference for the actual size of the character (i.e., the tall and tiny genie depicted in Fig. 5a and Fig. 5b, respectively). A prop was used to indicate the starting position of the hell dog (Fig. 5c). The movements of the dog are indicated with laser pointing, which helps the participants to imagine its position during the animation, giving them clues about the direction to look. Another physical prop (a hand sized ball) was used to provide a reference for the genie transformation and the movements of the fluctuating objects appearing at the end of the scene. During the experiment, the actor operating the laser pointing and props was trained to replicate their movement and position as identical as possible for all the trials, aided by some visual landmarks placed in the room. Even though copying the exact movements all the times is impossible for a human being, this limitation should be considered more as intrinsic to the TR approach rather than to the experiment design. Other stage props such as the magic spell book and the knife (Fig. 5d) are leveraged in both the methods, as they represent physical scene objects.

AR rehearsal: the participants were allowed to see through the OST-HMD all the virtual elements in the scene (Fig. 4b), i.e., the hell dog, the fluctuating objects, the genie (who was superimposed on the body of the actor controlling him via mocap), etc.

4.1.1 Procedure

Similarly to [2], the experiment was arranged with a mixed design (Fig. 6). The first part of the study followed a between-subjects design, by randomly assigning the participants to two equal-sized groups, each corresponding to a given rehearsal method (TR or AR). At first, the participants were introduced to the experiment and the script to be performed was presented. The participants were allowed to study and practice the lines of the script on their own for 15 minutes. Afterwards, they were requested to fill in a before-experience questionnaire (BEQ) concerning general information and demographics, previous experience with acting and with AR. Subsequently, the participants were asked to rehearse the scene three times using only the assigned method. Then, they underwent the actual scene shooting during which objective measures were gathered. It should be noted that, in order to perform a fair comparison, during the shooting the participants of both the groups acted without any visual aid (no AR contents and no laser pointing/props). At the end of the shooting, the participants were requested to fill in an after-shooting questionnaire (ASQ) to collect subjective feedback about the experience.

Lastly, with the aim to obtain a direct comparison of the two methods, a within-subject design was followed for the latter part of the experiment. The participants were requested to rehearse again the scene once by using the alternative method, and to fill in an after-alternative questionnaire (AAQ) to share their experience with the two alternatives.

4.1.2 Evaluation Criteria and Metrics

As said, in order to evaluate the two approaches, both objective and subjective measurements were collected and analyzed.

Subjective measurements: the three questionnaires used for the subjective evaluation (BEQ, ASQ and AAQ) are available at `http://tiny.cc/armocap_questionnaire`.

The ASQ, filled in after completing the shooting of the scene, was devised to investigate four factors: usability of the rehearsal method, perceived effectiveness, spatial and social presence. To measure the usability and learnability of the rehearsal approach, which are two factors deemed as important for the potential adoption of the system, the System Usability Scale (SUS) was adopted from [3]. The effectiveness of the rehearsal method was instead measured using statements adapted from [2], which the participants were asked to score on a 1-to-7 Likert scale (from completely disagree to completely agree). The sub-components of the effectiveness encompass all the aspects considered as relevant for acting in the selected use case, i.e., use of the acting space (positioning), gaze and eye following, gestures and proxemics, emotional expressiveness and engagement, and confidence in acting. With respect to [2], in this work spatial and social presence were additionally investigated using statements adapted from [17], with the aim to gain a more in-depth picture of these two factors. Statements were scored on a 1-to-7 Likert scale (from not at all to very much), and encompassed the following aspects: spatial awareness, interaction, placement of the scene elements and characters, sourcing of sounds, and reciprocating/reacting to the other characters' behaviour.

The aim of the AAQ, administered after the participants had experienced also the alternative rehearsal method, was twofold. On the one side, compare the two methods by asking the participants to indicate their preference for one of the methods. Statements in [2] were used, touching aspects like staging and positioning in the scene, awareness and control of the gazing direction, and synchronization when acting with the other characters. On the other side, evaluate the suitability of the AR method for supporting the rehearsal based on the approach exploited in [22]. Statements pertained the naturalness of interaction with the digital contents and the cognitive workload required to imagine the acted scene.

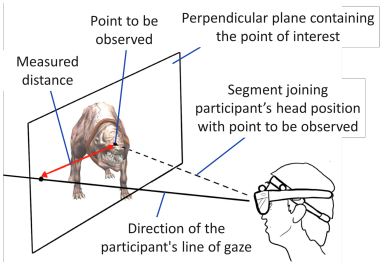


Figure 7: Computation of the eye distance metric.

Objective measurements: were harvested during the shooting for both the groups. Hence, also the participants in the TR group were requested to wear the OST-HMD while rehearsing the scene. The rationale behind this choice was to let them get accustomed in acting with the HMD (thus avoiding potential biases at evaluation time), and to provide them the audio cues for the acting. In this case, the functioning mode described in Section 3.2, i.e., with projection of AR contents disabled, was used. Specifically, for the TR group the audio was played back on the external speaker system (both at rehearsal and shooting time), whereas the AR group experienced the shooting with the audio played via the OST-HMD.

Objective measurements concerned two metrics: i) eye gaze and ii) spatial positioning. The first metric, in the following referred to as *eye distance*, was computed by measuring the distance between the point of interest and the point created by the intersection of the participant's gaze with the perpendicular plane containing the point of interest (as depicted in Fig.7). The metric was calculated at specific moments in time (collecting average data in a time-frame window centered at the event occurrence), when the script requested the participant to look at: the hell dog i) at the beginning of the scene ($EyeDist_{D1}$) and ii) while it is moving towards the participant ($EyeDist_{D2}$), iii) the genie's eyes while he is transforming ($EyeDist_{G1}$), and the two fluctuating objects at the end of the script iv) when they are appearing ($EyeDist_{F1}$) and v) when one of them has to be chosen ($EyeDist_{F2}$). The second metric, referred to as *hand distance*, evaluated the distance between the position in which the participant should position his or her hands and where he or she actually positioned them. The hand distance metric was calculated at the moment when the participant is requested to choose one of the fluctuating objects ($HandDist$). Data were collected by using the position/rotation of the HMD for the eye distance, and a glove tracked with the Optitrack system for the hand distance.

4.1.3 Sample

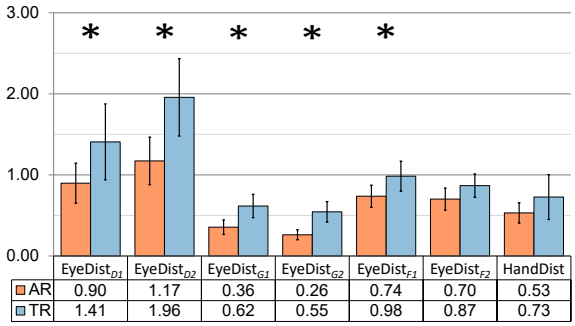
The user study was carried out involving 24 participants. The participants (15 males and 9 females) were aged between 21 and 45 ($\bar{x} = 25.98$, $sd = 4.25$). Regarding their acting expertise, most of them had good (50%) or some (29%) expertise. The remaining had low (17%) or no (4%) expertise. None of them had ever acted in scenes requiring mocap.

4.2 Results

In the following, the results obtained by measuring the above metrics are presented and discussed, with the aim to compare the performance of the TR and AR methods. Collected data were analyzed using MS Excel with the Real-Statistics add-on (v7.3). Statistical significance was evaluated using the two-tailed Student's t-test. Normality of data was verified with the D'Agostino-Pearson test.

4.2.1 Objective Results

For what it concerns the eye distance metric collected at shooting time, from Fig. 8 it can be observed that, in general, values were



	$EyeDist_{D1}$	$EyeDist_{D2}$	$EyeDist_{G1}$	$EyeDist_{G2}$	$EyeDist_{F1}$	$EyeDist_{F2}$	HandDist
AR	0.90	1.17	0.36	0.26	0.74	0.70	0.53
TR	1.41	1.96	0.62	0.55	0.98	0.87	0.73

Figure 8: Objective results based on the eye and hand distance metrics. Significant differences are marked with *.

smaller when the participants performed the rehearsal with the AR method. Differently than in [2], where Bouville et al. did not found any difference between TR and VR, in this work a statistically significant difference was spot for five out of the six events between TR and AR (in favor of AR). The differences were found for the parts of the script in which the participants had to look at the hell dog at the beginning of the scene (0.90 vs 1.41, $p = .008$), or while the dog is moving towards them (1.17 vs 1.96, $p = .032$). The other moments refer to situations in which the participants had to look at the genie's eyes during his transformation (0.36 vs 0.62, $p = .001$), interact with him after the resizing (0.26 vs 0.55, $p < .001$), and follow with the gaze the movements of the fluctuating objects (0.74 vs 0.98, $p = .009$). It is worth noticing that significant differences refer to events that correspond to different types of actions. For instance, $EyeDist_{D1}$ represents an emotional reaction that the participants had to simulate after seeing the hell dog. $EyeDist_{D2}$, $EyeDist_{G1}$ and $EyeDist_{F1}$ correspond to animations of virtual characters and objects that the participants had to follow with their eyes and react to. Finally, $EyeDist_{G2}$ refers to an interaction with a virtual character animated through mocap, whose size is significantly smaller than the participants' physiognomy.

Regarding the hand distance metric, no statistically significant differences were found.

4.2.2 Subjective Results

Based on collected results, the participants found the AR method as characterized by higher usability than the TR method (86.04 vs. 62.95, $p = .002$). According to the categorization given by the authors of [1], the score obtained by the AR method corresponds to a B grade, associated to the "Excellent" class in the adjective rating scale, whereas the TR was evaluated with a D grade, corresponding to the "Ok" class.

The reasons behind the higher appreciation for the AR method could be found by analyzing in detail the sub-scales of the SUS, which are reported in Table 2. The participants stated that they would be interested in using the AR method more frequently than the TR one (4.25 vs 3.46, $p = .043$), and judged the former as characterized by a lower complexity (1.25 vs 2.46, $p = 0.011$) and as easier to use (4.58 vs 3.46, $p = .015$) than the latter. The participants found the functionalities offered by the AR method more integrated in the system (4.83 vs 3.69, $p = .006$) and that the method was characterized by a lower inconsistency (1.33 vs 2.54, $p = .005$) compared to the TR one. Another interesting aspect regards learnability, as the participants felt that a lower amount of information had to be learned for using the AR method than the TR one (1.17, vs 2.08, $p = .025$).

Results in Fig. 9 show that the AR method scored better than the TR one for both spatial (5.56 vs 3.39, $p < .001$) and social presence (5.85 vs 4.13, $p = .001$). More specifically, regarding

Table 2: Subjective results concerning usability based on SUS [3]. Cells with a grey background highlight the best value (significant difference) for the two rehearsal methods.

Statement	AR	TR	*p*
I think that I would like to use this system frequently	4.25	3.46	.043
I found the system unnecessarily complex	1.25	2.46	.011
I thought the system was easy to use	4.58	3.46	.015
I think that I would need the support of a technical person to be able to use this system	2.25	2.38	.776
I found the various functions in this system were well integrated	4.83	3.69	.006
I thought there was too much inconsistency in this system	1.33	2.54	.005
I would imagine that most people would learn to use this system very quickly	4.08	3.54	.159
I found the system very cumbersome to use	1.75	2.62	.112
I felt very confident using the system	4.42	3.46	.055
I needed to learn a lot of things before I could get going with this system	1.17	2.08	.025
SUS Score	86.04	62.95	.002
Grade	B	D	
Adjective rating	Excell.	Ok	

spatial presence, the participants had a higher feeling that elements saw/heard/imagined were part of the environment in which they were located (5.67 vs 3.54, $p = .001$) and that they could be reached and touched (5.50 vs 3.54, $p < .001$) better with the AR method than the TR one. Moreover, the elements coming towards the participants (e.g., the hell dog) provoked more their instinctive reaction when using the AR method than the TR one (5.58 vs 3.54, $p = .003$). The AR method scored better than the TR one also for what it concerns the feeling of being part of the intended environment (5.58 vs 3.38, $p = .001$) and the directionality of sounds coming from seen/imagined objects (6.33 vs 3.77, $p < .001$). Finally, with the AR method the participants had more the instinct to touch objects they were seeing/imagining even though that action was not explicitly requested by the script (4.67 vs 3.23, $p = .024$). Regarding social presence, the participants stated that using the AR method they had a better feeling that the virtual character (i.e., the genie) was also able to see and hear them (5.25 vs 3.31, $p = .004$), letting them interact in a better and natural way with it (5.42 vs 3.77, $p = .002$). Moreover, the movements that, according to the script, the participants had to perform in response to events or to actions of the other virtual characters (e.g., the backward movement when the hell dog moved towards them) were easier to perform after the rehearsal with the AR method than the TR one (5.83 vs 4.08, $p = .006$). Finally, with the AR method, the participants felt to be more in the same environment of the other virtual characters (6.00 vs 4.46, $p = .004$), which allowed them to make an easier eye contact (6.42 vs 3.77, $p < .001$) or have more natural interactions (6.00 vs 4.77, $p = .002$) than with the TR one.

Considering the overall effectiveness of the rehearsal method, the participants judged the AR method as more helpful than the TR one to shoot the scene after the three rehearsals without any clue about the virtual elements in the scene (6.43 vs 4.57, $p < .001$). In particular, the AR method helped the participants to better position in the environment (6.92 vs 5.38, $p < .001$), improve to use the space (6.92 vs 5.15, $p < .001$), follow the movements of the other characters (6.33 vs 4.54, $p = .001$), and feel comfortable with their gestures while shooting (6.50 vs 4.54, $p = .001$) than the TR one. Moreover, the AR method allowed the participants to better express the emotional states of the character they were playing with facial expressions (6.00 vs 3.62, $p < .001$) and gestures (6.08 vs 3.85, $p < .001$) than the TR one. The AR method also helped the participants to do better on their emotional involvement (6.25 vs 4.31, $p < .001$).

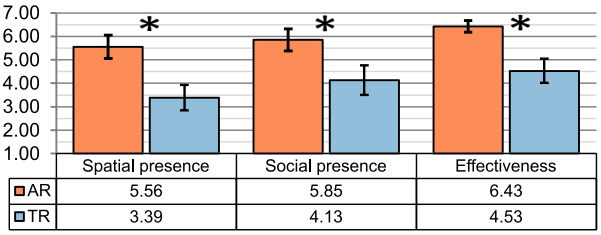


Figure 9: Subjective results concerning spatial presence, social presence, and effectiveness as investigated in [2, 17]. Significant differences are marked with *.

Table 3: Subjective results concerning the suitability of the AR method based on the analysis tool proposed in [22].

Statement	$\bar{x}$ ($c.i._{95\%}$)
Watching the virtual objects was as natural as watching real-world objects	5.25 (0.55)
I had the impression that virtual and real objects belonged to the same world	4.58 (0.73)
I had the impression that I could touch and grasp the virtual objects	5.54 (0.56)
I had the impression that the virtual objects were in the real world rather than simply projected on a screen	5.00 (0.76)
I had the impression of seeing virtual objects as three-dimensional and not as mere flat images	6.46 (0.35)
I do not notice differences between real and virtual objects	4.29 (0.62)
I had not to make an effort to recognize virtual objects as being three-dimensional	6.25 (0.36)

Finally, compared to the TR method, the AR method allowed the participants to be more confident about their performance while shooting (6.58 vs 4.85, $p < .001$), thus making them more ready to shoot (6.25 vs 4.92, $p = .001$).

From a statistical standpoint, overall the results showed a difference between TR and AR larger than that observed between TR and VR by the authors of [2]. Analogously, in the study of Kammerlander et al. [12], a significant difference for the social presence factor was not found. These findinds about effectiveness and social presence may be ascribable to a difference in the sample characteristics, being the participants of the current study less skilled in mocap acting with respect to the ones in [2, 12], or to a higher effectiveness/ability to stimulate social presence of AR against VR for the considered scenario. However, further investigations directly comparing the two technologies (AR and VR) shall be performed in the future to clarify this aspect. Despite the above differences, similar trends were observed (AR/VR scoring better than TR), with the exception of the positioning in the scene, for which TR proved to work better.

For what it concerns the direct comparison of the TR and AR methods, the obtained results confirmed the higher appreciation for the AR method compared to the TR one. More specifically, the participants indicated the AR method as preferable with respect to TR one for positioning in the scene (91.67% vs 8.33%), controlling gaze direction (95.83% vs 4.17%), synchronizing with the other characters (91.67% vs 8.33%), and eliciting emotional involvement (95.83% vs 4.17%). Overall, 100% of the participants preferred the AR method to the TR one.

Finally, the soundness of the AR method is further substantiated by the relatively high scores assigned by the participants to statements concerning its suitability, reported in Table 3.

5 Conclusions and Future Work

The goal of this work was to investigate the use of AR to help actors in the rehearsal of mocap scenes. A system named AR-MoCap is proposed, letting actors wearing an OST-HMD visualize in real time virtual characters superimposed on the actors who are controlling them through mocap, together with both real and computer-generated elements (virtual objects, animations, and visual effects). The proposed system has been implemented using HoloLens 1st Gen. as OST-HMD, and Optitrack as optical tracking system for mocap.

An experimental evaluation were carried out with the aim to assess the effectiveness of the proposed AR rehearsal method, comparing it with the traditional approach based on physical props and laser pointing in both objective and subjective terms.

According to the obtained results, the AR method scored better than the TR one in terms of usability, spatial presence, social presence, and perceived effectiveness. Moreover, the results indicated that the AR method can be particularly effective for actors to train in directing their gaze towards virtual elements in the scene. This is especially true when they have to interact with virtual objects that move, like animated characters (also with mocap), or when they have to react emotionally to a virtual event. These aspects could not be observed in works, such as [5, 6, 10, 16], since they did not consider interactions among multiple users/actors.

It is worth noticing that, differently than the scenario in [8], that considered characters of the same size of the actor and not controlled with mocap, the results were obtained considering mocap-animated characters characterized by varying scales (different than those of the human actor controlling them), thus proving that the proposed system could be helpful to rehearse also this kind of scenes.

Several ways for extending the experimental analysis reported in this paper can be envisaged. First, even though the proposed AR-MoCap system supports two different use cases (i.e., augmenting the other characters in the scene, and augmenting the own character), due to technological limitations of the OST-HMD the experiment focused just on one of them; thus, in the future, the system shall be evaluated also with the other use case. Second, further experiments could be carried out by involving more than one actor controlling virtual characters through mocap. Third, other technologies adopted in virtual production like, e.g., VR or LED walls could be compared. In all the cases, the evaluation should be widened to include also actors with previous experience in mocap.

The AR-MoCap system could be exteded as well. For instance, the possibility to support more than two actors simultaneously wearing an OST-HMD could be added. Moreover, adopting a marker-less tracking system would allow the actors to move in larger spaces without the need to rely on potentially intrusive equipment such as Optitrack's reflective markers. Tracking systems based, e.g., on inertial sensors, could be used for this purpose, since the MotionHub software used for the implementation of AR-MoCap already enables communications and interoperability with other mocap systems; this integration has been already explored, but it was found that the usage of such marker-less tracking system introduces new issues (e.g., latency, tracking inaccuracies, etc.) that would have to be addressed. Another possibility could be to implement mocap by leveraging the images captured by the cameras embedded in the HMD; although this solution could allow to extend the tracking area, it would still present challenges related to occlusion problems and computation/network load on the HMD. Finally, another possible extension could concern the integration of facial tracking; in this way, the actors wearing an HMD would be allowed to visualize not only the articulated bodies of the virtual characters but also their facial expressions.

Acknowledgments

This work has been developed in the frame of the VR@POLITO initiative. The research was supported by PON "Ricerca e Innovazione" 2014-2020 – DM 1062/2021 funds.

REFERENCES

[1] B. Aaron, K. Philip, and M. James. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[2] R. Bouville, V. Gouranton, and B. Arnaldi. Virtual reality rehearsals for acting with visual effects. In *International Conference on Computer Graphics & Interactive Techniques*, pp. 1–8, 2016.

[3] J. Brooke et al. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[4] A. Cannavò, F. G. Pratticò, G. Ministeri, and F. Lamberti. A movement analysis system based on immersive virtual reality and wearable technology for sport training. In *Proceedings of the 4th international conference on Virtual Reality*, pp. 26–31, 2018.

[5] X. Chen, Z. Chen, Y. Li, T. He, J. Hou, S. Liu, and Y. He. Immertai: Immersive motion learning in VR environments. *Journal of Visual Communication and Image Representation*, 58:416–427, 2019.

[6] I. Damian, M. Obaid, F. Kistler, and E. André. Augmented reality using a 3D motion capturing suit. In *Proceedings of the 4th Augmented Human International Conference*, pp. 233–234, 2013.

[7] R. Ge and T.-C. Hsiao. A summary of virtual reality, augmented reality and mixed reality technologies in film and television creative industries. In *IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, pp. 108–111. IEEE, 2020.

[8] R. Ichikari, R. Tenmoku, F. Shibata, T. Ohshima, and H. Tamura. Mixed Reality pre-visualization for filmmaking: On-set camera-work authoring and action rehearsal. *The International Journal of Virtual Reality*, 7(4):25–32, 2008.

[9] A. Ikeda, D.-H. Hwang, H. Koike, G. Bruder, S. Yoshimoto, and S. Cobb. AR based self-sports learning system using decayed dynamic timewarping algorithm. In *ICAT-EGVE*, pp. 171–174, 2018.

[10] S. Ikeda, T. Taketomi, B. Okumura, T. Sato, M. Kanbara, N. Yokoya, and K. Chihara. Real-time outdoor pre-visualization method for videographers—real-time geometric registration using point-based model. In *IEEE International Conference on Multimedia and Expo*, pp. 949–952. IEEE, 2008.

[11] N. Kadner. The virtual production field guide. *Epic Games*, 2019.

[12] R. K. Kammerlander, A. Pereira, and S. Alexanderson. Using virtual reality to support acting in motion capture with differently scaled characters. In *IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 402–410. IEEE, 2021.

[13] A. Kumar. Introduction to visual effects (VFX). In *Beginning VFX with Autodesk Maya*, pp. 1–10. Springer, 2022.

[14] P. Ladwig, K. Evers, E. J. Jansen, B. Fischer, D. Nowottnik, and C. Geiger. MotionHub: Middleware for unification of multiple body tracking systems. In *Proceedings of the 7th International Conference on Movement and Computing*, pp. 1–8, 2020.

[15] C. Ling and W. Zhang. ARFMS: An AR-based WYSIWYG filmmaking system. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 12(6):4345–4352, 2014.

[16] J. Liu, Y. Zheng, K. Wang, Y. Bian, W. Gai, and D. Gao. A real-time interactive tai chi learning system based on VR and motion capture technology. *Procedia Computer Science*, 174:712–719, 2020.

[17] M. Lombard, T. B. Ditton, and L. Weinstein. Measuring presence: The temple presence inventory. In *Proceedings of the 12th annual international workshop on presence*, pp. 1–15, 2009.

[18] A. Menache. *Understanding motion capture for computer animation and video games*. Morgan kaufmann, 2000.

[19] S. G. M. Nassar. Engaging by design: Utilization of VR interactive design tool in mise-en-scène design in filmmaking. *International Design Journal*, 11(6):65–71, 2021.

[20] J. A. Okun, V. Susan Zwerman, et al. *The VES handbook of visual effects: industry standard VFX practices and procedures*. Routledge, 2020.

[21] F. Pires, R. Silva, and R. Raposo. A survey on virtual production and the future of compositing technologies. *AVANCA— CINEMA*, pp. 692–699, 2022.

[22] H. Regenbrecht and T. Schubert. Measuring presence in augmented reality environments: Design and a first test of a questionnaire. *arXiv preprint arXiv:2103.02831*, 2021.

[23] S. Rizvic, V. Okanovic, and D. Boskovic. Digital storytelling. In *Visual computing for cultural heritage*, pp. 347–367. Springer, 2020.

[24] A. Sanna, F. Lamberti, F. De Pace, R. Iacoviello, and P. Sunna. ARSSET: Augmented reality support on SET. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, pp. 356–376. Springer, 2017.

[25] M. Seymour. Art of led wall virtual production, part one: Lessons from the mandalorian. *Retrieved from fxguide. com: https://www. luxmc. com/press-a/art-of-led-wall-virtual-production-part-one-lessons-from-the-mandalorian/. Accessed*, 30, 2022.

[26] T. Soghomonian. Ian McKellen: Filming "The Hobbit" made me cry with frustration'. `https://www.nme.com/news/film/ian-mckellen-filming-the-hobbit-made-\cry-with-f-877575`, 2012. (Accessed 10/07/2022).

[27] A. Stamm, P. Teall, and G. B. Benedicto. Augmented virtuality in real time for pre-visualization in film. In *IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 183–186. IEEE, 2016.

[28] H. Tamura, T. Matsuyama, N. Yokoya, R. Ichikari, S. Nobuhara, and T. Sato. Computer vision technology applied to mr-based pre-visualization in filmmaking. In *Asian Conference on Computer Vision*, pp. 1–10. Springer, 2010.

[29] A. Whitehurst. The visual effects pipeline. `http://www.andrew-whitehurst.net/pipeline.html`, 2008. (Accessed 10/07/2022).