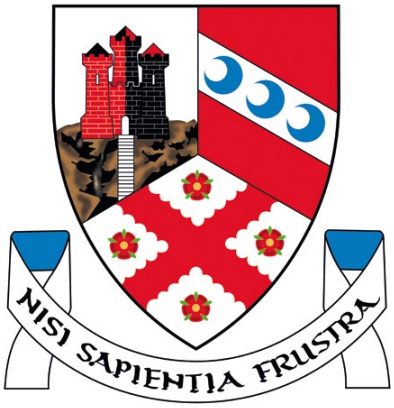
Intermediated Reality



Llogari Casas Cambra

School of Computing

Edinburgh Napier University

A thesis submitted in partial fulfilment of the requirements of Edinburgh Napier University, for the award of

Doctor of Philosophy

Distro ITN May 2020

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

External examiner Prof. Anthony Steed

Internal examiner Dr. Augusto Abreu Esteves

Director of studies Prof. Kenny Mitchell

Additional supervisors Dr. Kevin Chalmers, Dr. Gergory Leplatre

Llogari Casas Cambra

May 2020

Abstract

Real-time solutions to reducing the gap between virtual and physical worlds for photo-realistic interactive Augmented Reality (AR) are presented. First, a method of texture deformation with image inpainting, provides a proof of concept to convincingly re-animate fixed physical objects through digital displays with seamless visual appearance. This, in combination with novel methods for image-based retargeting of real shadows to deformed virtual poses and environment illumination estimation using inconspicuous flat Fresnel lenses, brings real-world props to life in compelling, practical ways.

Live AR animation capability provides the key basis for interactive facial performance capture driven deformation of real-world physical facial props. Therefore, *Intermediated Reality* (IR) is enabled; a tele-present AR framework that drives mediated communication and collaboration for multiple users through the remote possession of toys brought to life.

This IR framework provides the foundation of prototype applications in physical avatar chat communication, stop-motion animation movie production, and immersive video games. Specifically, a new approach to reduce the number of physical configurations needed for a stop-motion animation movie by generating the in-between frames digitally in AR is demonstrated. AR-generated frames preserve its natural appearance and achieve smooth transitions between real-world keyframes and digitally generated in-betweens. Finally, the methods integrate across the entire Reality-Virtuality Continuum to target new game experiences called Multi-Reality games. This gaming experience makes an evolutionary step toward the convergence of real and virtual game characters for visceral digital experiences.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Kenny Mitchell for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my dissertation committee members (Dr. Gregory Leplatre and Dr. Kevin Chalmers), my panel chairs (Prof. Ben Paechter and Prof. David Benyon), my external examiner (Prof. Anthony Steed) and my internal examiner (Dr. Augusto Abreu Esteves) for their great support and invaluable advice.

Further, I would like to thank my Napier and DISTRO project colleagues for their continued support over the course of my Ph.D. studies. Thanks are also due to the European Commission. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant 642841.

Last but not least, I would like to express my deepest gratitude to my family and friends. This dissertation would not have been possible without their warm love, continued patience, and endless support.

Vull dedicar aquesta tesi a la meva estimada germana, Laura, als meus pares, Mercè i Llogari, i als meus avis, Montse, Joan, Emilia i Manel.

Sa aking minamahal na Ampie.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The concept of enhancing the real-world with overlaid content has been present since the very beginning of the 16th century. *"Magia Naturalis"*, from [Giambattista della Porta, 1584], is the first documented reference conceptualizing this idea with the *Pepper's ghost* illusion technique. Early in the 20th century, [L. Frank Baum, 1901] envisioned the use of electronic displays to overlay characters into the real-world through his novel, *"The Master Key"*. He introduced an approximation of what it is currently known as *Augmented Reality (AR)*. This technology uses the capabilities of a computer generated display to enhance the user's real-world experience. It belongs to a set of environments defined as *Mixed Reality (MR)* experiences that aim to present the real and virtual world unified in the same space and time. This term was first defined by [Milgram et al., 1994] in the *Reality-Virtuality Continuum (RVC)*.

Mixed Reality (MR) experiences start at the real environment itself and cover technologies such as *Augmented Reality (AR)*, *Augmented Virtuality (AV)* and *Virtual Reality (VR)* (see figure 1.1). These environments can either be experienced by sole or multiple participants at the same time. When multiple users interact through a MR environment, these experiences got defined as *Collaborative Mixed Reality (CMR)* by [Billinghurst and Kato, 1999].

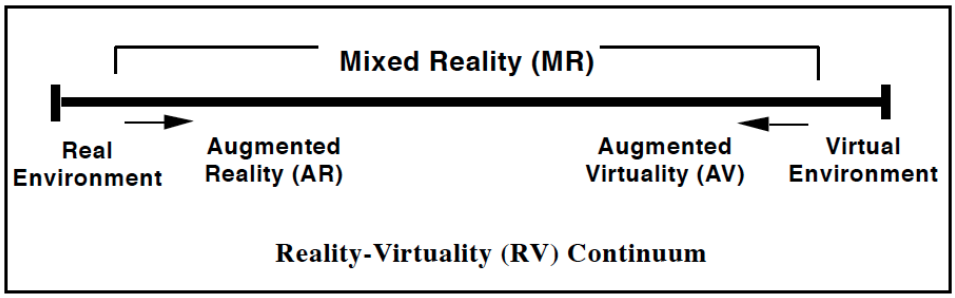


Fig. 1.1 *Reality-Virtuality Continuum (RVC)* [Milgram et al., 1994].

Intermediated Reality (IR) addresses the gap between virtual and physical worlds using photo-realistic interactive AR to convincingly re-animate physical objects through digital displays and proposes a CMR framework to stimulate mediated collaboration among distributed peers. By altering the camera video feed with a reconstructed appearance of the object in a deformed pose, we perform the illusion of movement in real-world objects to realize collaborative tele-present AR.

Our framework aims not only to allow users to collaborate remotely in a novel way, but also to enhance creativity, imagination and interaction with inanimate objects of our daily lives. In this sense, an AR system capable of animating real world objects and toy figurines is proposed. The research presented in this thesis allows participants to collaboratively interact with each other using inanimate objects and toys as if they were brought to life. This is done through a framework that traverses the RVC.

1.1 Research Questions

The following research questions address the gap between virtual and physical worlds using interactive AR to convincingly re-animate physical objects through digital displays.

- **RQ1**. *How are physical objects convincingly re-animated through digital displays using mobile AR?*

  - **RQ1.1**. *How are areas not present in the physical object reconstructed when revealed in AR?*

- **RQ2**. *How are physical shadows plausibly retargeted according to the animated movement of the real-world object in AR?*

  - **RQ2.1**. *How is the environmental illumination estimated employing a portable approach that seamlessly blends with everyday objects?*

- **RQ3**. *How are objects brought to life integrated into a collaborative distributed environment?*

1.2 Contributions

This thesis makes a number of contributions to the field of tele-present AR by addressing the gap between virtual and physical worlds to convincingly re-animate physical objects through digital displays.

- we introduce *Deformable Object Retargeting*. This method achieves an illusion of movement from the real-world object through image retargeting techniques using AR (see section 3.4, [Casas et al., 2017]).

- we introduce appearance reconstruction, a method that is capable of reconstructing areas not present in the physical reference model when revealed in AR (see section 3.4.2, [Casas and Mitchell, 2019]).

- we present *Shadow Retargeting*. This method is an efficient and focused approach in which already present shadows from static real objects in the scene are retargeted according to virtual overlaid AR movement (see section 3.5, [Casas et al., 2018b]).

- we introduce a novel light-estimation approach employing a flat light probe fabricated using a flat reflective Fresnel lens. Unlike three-dimensional sampling probes, such as reflective spheres, Fresnel lenses can be incorporated into any type of object, such as a book or product packaging, providing a seamless light-estimation approach for AR experiences (see section 3.3, [Casas et al., 2019]).

- we introduce, *Intermediated Reality (IR)*, a tele-present AR framework that enables mediated communication and collaboration for multiple users through the remote possession of toys brought to life (see chapter 4, [Casas and Mitchell, 2019]).

- we showcase applications of IR is for distributed communication and collaboration. Further, we present *Multi-Reality Games*, an innovative form of gaming that encompasses interactions with real and virtual objects throughout the spectrum of the RVC (see chapter 5, [Casas et al., 2018a] [Casas and Mitchell, 2019]).

1.3 Publications

[Casas et al., 2017]. Casas, L., Kosek, M., & Mitchell, K. (in press). Props Alive: A Framework for Augmented Reality Stop Motion Animation. *In 2017 IEEE 10th Workshop on Software Engineering and Architectures for Realtime Interactive Systems.*

[Casas et al., 2018a]. Casas, L.*, Ciccone, L.*, Cimen, G., Wiedemann, P., Fauconneau M., Mitchell K. & Sumner B. Multi-Reality Games: an Experience Across the Entire Reality-Virtuality Continuum. *ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI 2018).*

[Casas et al., 2018b]. Casas, L., Fauconneau, M., Kosek, M., McLister, K., & Mitchell, K. Image Based Proximate Shadow Retargeting. *In Computer Graphics and Visual Computing (CGVC) 2018.*

[Casas et al., 2019]. Casas, L., Fauconneau, M., Kosek, M., McLister, K., & Mitchell, K. Enhanced Shadow Retargeting with Light-Source Estimation Using Flat Fresnel Lenses. *MDPI Journal in Computers - Special Issue "Selected Papers from the Computer Graphics & Visual Computing (CGVC 2018).*

[Casas and Mitchell, 2019]. Casas, L., & Mitchell, K. Intermediated Reality: A Framework for Communication through Tele-Puppetry. *Frontiers in Robotics and Artificial Intelligence - Special issue "Collaboration in Mixed-Reality" 2019.*

[Casas, 2019]. Casas, L. Intermediated Reality. *SA '19: SIGGRAPH Asia 2019 Doctoral Consortium.*

1.4 Thesis Overview

The remainder of this document is structured as follows:

- Chapter 2 discusses the current state of the field and helps further define the scope of the research by grounding it in the literature.
- Chapter 3 introduces *Object* and *Shadow Retargeting* with appearance reconstruction. These methods allow static real-world objects to be animated seamlessly in AR as if they were brought to life.
- Chapter 4 describes Intermediated Reality and introduces *ToyMeet*, a CMR system that uses toys brought to life for distributed communication.
- Chapter 5 introduces IR as a communication tool and showcases an industrial application for stop motion animation. Further, it presents *Multi-Reality Games*, an innovative form of gaming that encompasses interactions across the RVC.

- Chapter 6 summarises the contributions of this thesis to the research field, provides answers to the posed research questions and outlines how the work presented here can be extended in the future.

Chapter 2

Related Work

Intermediated Reality relates to a variety of research areas in AR to integrate a CMR system. Our framework bases its structure on related works in MR and graphical user interfaces (section 2.1). IR relates to physical avatars and robots to support remote collaboration in a mediated communication environment (section 2.2). We build upon previous experimental systems to develop a novel framework that enables bringing props to live (section 2.3). Concretely, this thesis draws from *Image Retargeting* techniques to animate physical puppets and shadows (section 2.4). Further, related applications for IR are described (section 2.5). Finally, section 2.6 summarises related work and highlights the contributions of this thesis to the research field.

2.1 Collaborative Mixed-Reality Systems

[Billinghurst and Kato, 1999] defined CMR systems as a natural medium for *Computer Supported Collaborative Work (CSCW)*. [Billinghurst et al., 1998] proposed an early example of CMR cooperation with face-to-face experiences using hand-held and *Head Mounted Displays (HMD)* in AR. [Kiyokawa et al., 2002] validated that cooperation through MR technologies could significantly improve collaboration within users by merging the real and

virtual world together. This shared MR context allows to use the same non-verbal cues used in face-to-face conversations, while also interacting with AR content overlaid in front of users. More recently, [Zünd et al., 2015] proposed a system in which multiple collaborators could simultaneously enhance creativity in AR. [Fairchild et al., 2017] and [Steed et al., 2012] embody processes of capturing remote participants beamed into the shared space with visual depictions of them displayed digitally in mixed reality. Our system provides a real-world physical intermediary for the remote person's presence with animated expression synchronized with transmitted vocal audio.

While research on CMR systems has focused primarily on real-time collaboration, less work has been done on the use of AR for asynchronous participation. [Renevier and Nigay, 2001] introduced an early CMR system that allowed the creation and visualization of AR messages in real-world space for archaeologists. Such system was extended by [Nassani et al., 2015] to allow users to place virtual labels on any object or location in the real world. [Kooper and MacIntyre, 2003]'s work is another pioneering example of an asynchronous CMR system. This research developed an AR browser that could get registered to a specific real-world location and became visible by other participants.

Related schemes for asynchronous messaging have been introduced by [Everitt et al., 2003] and [Kjeldskov et al., 2009] using interactive boards for enhanced communication between team members. These systems allow users to leave non synchronized messages to other participants. Our research builds on these foundations to create a turn-based interactive avatar chat for tele-puppetry in a tele-present *Intermediated Reality* system. Each participant first sees his own vocalization and facial expressions captured locally, then transmits the message to a database and is finally made available to the receiver's physical puppet when the user accesses the chat (see chapter 4).

2.2 Physical Avatars for Remote Collaboration

The use of physical objects as avatars for remote collaboration has been conceived previously. For instance, [Sekiguchi et al., 2001] used *Robotic User Interfaces (RUI)* to communicate shapes and movements with each other using snake-like robots equipped with servomotors that responded to information transmitted through a network. Another example proposed by [Sekiguchi et al., 2004] is *RobotPHONE*, a RUI system that enabled users to communicate shapes and motions with each other using physical toys robots. We base our approach on the same foundations, triggering the AR animations on real-world according to the information received from the remote participant.

[Yim and Shaw, 2011] proposed interactive bidirectional robot intermediaries for performing tasks and applications. Our approach draws from this concept to use toys figurines as interlocutors in both ends of a remote communication.

Drawing from *ClayVision* by [Takeuchi and Perlin, 2012], a framework capable of photo-realistically animate city buildings through AR, this thesis aims to create the illusion of movement on static puppets by seamlessly rendering synthetic objects into a real-world scene. Our system brings the advantages of both distributed and mediated avatars, to propose an AR system capable of animating real world objects and toy figures with photo-realistic results. The research presented in this thesis allows participants to interact with inanimate objects and toys as if they were brought to life (see chapter 4).

2.3 Experimental Framework

Over the literature course, software design frameworks have been developed and applied to a large variety of use-cases in computer graphics. [Kennedy et al., 1996] introduced a novel object oriented design framework in information visualization that mapped different development platforms and use cases. This framework unified models of *Model-View-*

Controller (MVC) with others, such as *Presentation-Abstraction-Control (PAC)* and *User Interface Management Systems (UIMS)*. [Bauer et al., 2001] proposed an AR component-based framework capable of being used for a variety of proposes by handling hardware and software separately. This structure set the design basis for module-based structures in AR. Their framework implemented three main core modules: *Application*, *Tracking* and *Interaction*.

Research by [Bauer et al., 2001] presented a solid approach to solve low-level interaction with the device hardware. This facilitated developing AR-based frameworks mapped to specific solutions. For instance, a framework for creating AR presentations was introduced by [Ledermann and Schmalstieg, 2005]. In this specific solution, a component-based framework was employed following [Bauer et al., 2001]'s approach. This was further extended to accommodate application-specific modules. In this thesis, we develop a software framework that implements each research contribution as an independent module (see section 3.2). This framework uses the foundations of *experimental design* with the aim to determine if formulated research questions in section 1.1 can consistently relate to acquired results.

2.4 Appearance Retargeting

Drawing from the research introduced by [Karsch et al., 2011], in which they achieved to render photo-realistic synthetic objects into legacy images, we aim to create the illusion of movement on static puppets by seamlessly rendering synthetic objects into a real-world scene. To create such appearance, we draw from *Image Retargeting* approaches. These techniques aim to retarget an image to a form with visual constraints while preserving its main features. Maintaining an aesthetically plausible output when performing deformations on an image can be a challenging problem to address. Approaches from [Setlur et al., 2005] and [Goferman et al., 2010] use content-aware importance maps and regions of interest to perceptually guide and steer image sampling in the retargeted reconstruction. In this thesis,

we apply the concept of image retargeting to AR using *Object Deformed Retargeting* (see section 3.4), in the case where we have known segmented 3D geometry of physical objects in the image and the ability to register the mesh pose in real-time with marker-less tracking introduced by [Comport et al., 2006].

2.4.1 Light estimation

[Knecht et al., 2011] observed that if photorealism wants to be achieved in AR, virtual meshes need to be adequately relighted according to the real-world environment. This problem introduced what is known as the relighting approach, in which an already illuminated virtual mesh needs to be correctly shaded according to real-world light sources.

[Ramamoorthi and Hanrahan, 2001] introduced an efficient approximate, but reasonably accurate, light representation using spherical harmonics. This early method used nine coefficient models for calculating its low-frequency spherical harmonic representations. A quadratic or higher order polynomial equation was employed to shade virtual meshes according to their surroundings. Subsequently, [Jung et al., 2007] extended this technique using precomputed radiance functions. Further, [Franke and Jung, 2008] enabled real-time relighting using environment maps under a pre-set of conditions. Currently, spherical harmonics are widely used in AR as they provide a compact mathematical model for the global lighting environment around a point in space, describing the distribution and colors of multiple directional light sources.

[Grosch, 2005] proposed a related technique to calculate lighting variance in the scene using differential photon mapping. Using this approach, virtual objects were relighted according to the shading difference between real and virtual scenes. [Grosch et al., 2007] further extended this approach using light fields to enable consistent augmentation of virtual objects with both direct and indirect lighting. [Knecht et al., 2010] added instant radiosity to his line of work to achieve a plausible MR environment in real-time.

More recent results approached light estimation with real-world scene helpers. These are predefined artefacts used for reflectance estimation. [Basri and Jacobs, 2001] proposed the use of a reflective chrome sphere to extract distant light sources using their surface normal vector. [Calian et al., 2013] used a 3D-printed shading probe to capture both hard and soft shadow behaviour from general illumination environments. [Knorr and Kurz, 2014] and [Calian et al., 2018] used human faces as light probes for coherent rendering of virtual content. However, face tracking inaccuracy is a limiting factor on the values estimated. This approach was limited to outdoor environments. A variant of light estimation from faces further developed in [ARKit, 2019]. [Weber et al., 2018] extended this line of work by providing a general solution for indoor and outdoor environments. [Boom et al., 2017] observed that image and depth textures can provide enough information to reconstruct light-source positions using shape from shading, given shadow observation from depth map geometry. [Mandl et al., 2017] enabled a related scheme for relighting using every-day objects using *Convolutional Neural Networks (CNN)*. This method used learned light probes over a pre-established object to detect environment lighting. Nonetheless, this technique requires to previously train a CNN with all possible types of lighting of an object, and to extract and store spherical harmonics for each view.

In this thesis, we introduce a light-estimation approach that enables light-source detection using flat Fresnel lenses embedded in every-day objects (see section 3.3). Drawing from the reflective chrome sphere method from [Basri and Jacobs, 2001], we estimate light-source positions through a 180 degree planar Fresnel film. We embed this reflector in detectable objects, enabling light-source estimation in mobile devices without additional gadgets.

2.4.2 Scene Relighting

Image Retargeting approaches for AR are relatively novel and the first method introduced by [Leão et al., 2011], *Altered Reality*, is a relatively isolated work. This method deformed

the appearance of a physical cube by projecting a virtual mesh on top of it and performing texture deformation from the cube. This demonstration was very limited to basic shapes and did not address indirect re-lighting in the form of shadow retargeting. Additionally, real-time performance was limited to high-end desktop computers. In contrast, our method introduces direct and indirect relighting on mobile devices in real-time.

[Takeuchi and Perlin, 2012] introduced *ClayVision*, a related scheme for building large scale objects in the city. This method pre-scanned a set of known locations and reconstructed them into meshes. Their camera feed was patched with a pre-modified version of the background image, in which the appropriate building was previously removed and manually inpainted. Upon the movement of the building's virtual animated mesh, the desired deformed pixels were overlaid on the camera image. While this achieved real-time performance in mobile devices, it did not address any means of indirect relighting. This thesis extends state of the art work with a dynamic texture lookup according to the overlaid animation over time for the purpose of shadow retargeting (see section 3.5).

2.4.3 Augmented-Reality Shadows

[Sato et al., 1999] introduced an early off-line method to recover scene illumination for AR using shadow brightness. [Haller et al., 2003] first provided a method to recreate real-time shadow volumes in a MR scenario. [Jacobs et al., 2005] further extended this by addressing multiple light sources and shadow overlapping. [Nowrouzezahrai et al., 2011] presented an approach that captured and factorized external lighting in a manner that allowed for realistic relighting of both animated and static virtual objects. [Castro et al., 2012] depicted an AR soft shadow estimation approach using a reflective light-probe sphere. This method only recovered a single light source and shadows cast by static objects. [Calian et al., 2013], using the *3D-printed Shading Probe*, captured shadow behaviour into a piecewise constant spherical harmonics basis representation that was directly applied to rendered diffuse global

illumination. The design of a single 3D-printed shading probe that captured both hard and soft shadow behaviour from general illumination environments proved to be complex.

Methods for removing shadows in computer graphics have been developed to alter the real-world. [Jaynes et al., 2001] introduced a technique that detected and corrected transient shadows in a multi-projector display. [Baba and Asada, 2003] proposed an off-line method to remove shadows from images based on color space analysis and showed experimental results of shadow removal and virtual shadow overlay. [Kakuta et al., 2008] described a method to detect moving objects and remove their shadows for superimposing them on MR systems. They provided an approach to cut out the foreground from a real image using a probability-based segmentation method and superimpose virtual objects using the stencil buffer. [Iwai et al., 2014] described a shadow removal technique for a multiple overlapping projection system. In particular, research situations where cameras could not be placed between the occluder and projection surface. [Zhang et al., 2015] presented a related case for shadow inpainting in which an illumination recovering optimization method was used for static shadow removal. Nonetheless, due to the need of high performance real-time graphics for mobile devices, our approach presented in this thesis is based on proximity coherence with direct sampling (see section 3.5.3).

In this thesis, we focus on *Shadow Retargeting* by performing shadow reconstruction to a pose different from its observed one (see section 3.5). We showcase our Shadow Retargeting approach in a variety of estimated real and accurate synthetic light-environment scenarios ranging from single-point light illumination to area sources and natural illumination environments.

2.4.4 Image Inpainting

As defined by [Mann, 2001], in an ideal MR environment, the user should be able to add and subtract visual content on equal basis. *Image inpainting*, also known as *Diminished Reality (DR)*, aims to subtract information from a live camera feed achieving plausible results.

Real-time scene reconstruction with structural and perceptual consistency has demonstrated to be computationally expensive and early inpainting examples from [Bertalmio et al., 2003] were limited to off-line techniques. *PixMix* by [Herling and Broll, 2012] achieved real-time inpainting with few background constraints and without the need of a multi-view approach. This algorithm iteratively minimized the spatial and appearance cost into a function achieving consistency over patterned backgrounds. Nevertheless, on-line performance was still limited to high end processors on portable computers.

The use of machine learning to solve the inpainting problem has been introduced recently. [Yeh et al., 2017] proposed a method to perform semantic image inpainting with deep generative models. [Yang et al., 2017] achieved high-resolution image inpainting using multi-scale neural patch synthesis. Nonetheless, these techniques are still limited to off-line performance.

In the case of mobile devices, methods often require additional information. [Jarusirisawad and Saito, 2007] introduced an approach that used multiple hand-held devices to perform background inpainting. [Kawai et al., 2013] proposed a scene reconstruction approach based on depth segmentation for non-planar backgrounds. In our work, we use a background observation method, similar to *ClayVision* from [Takeuchi and Perlin, 2012], in which the user captures the background of the scene before initializing IR (see section 3.5.12).

2.5 Applications

Section 2.5.1 describes related experiences across the RVC that are alike to *Multi-Reality Games* in IR. Further, a flexible and optimized tool for Stop Motion animation artists is presented in this thesis. Section 2.5.4 compares related work in the field with our approach.

2.5.1 Mixed Reality experiences

During the last decade of the 20th century, MR experiences had an exponential growth due to the advancement of computing capabilities and resources. Research performed on the field of AR got divided into three main subcategories: Head-Mounted Augmented Reality, Spatial Augmented Reality and Hand-held Augmented Reality.

Head-Mounted Augmented Reality

HMDs were present since the very beginning of MR technologies. The first HMD connected to a computer was invented by [Sutherland, 1968] and was named *"The Sword of Damocles"*, due to its heaviness and uncomfortableness, as an allusion to the classical Greek mythology of homonymous name. [Rosenberg, 1992] created the first functioning HMD AR system, entitled *"Virtual Fixtures"*, proving objectively to be beneficial for trainees. [Feiner et al., 1993] presented the first research paper featuring an AR prototype entitled *"KARMA"*, an AR tool for maintenance assistance.

In Head-Mounted AR, computer-generated objects tend to be displayed on the screen as holograms. [Gabor, 1948] was the inventor of holography, a technique that refracts light wavelengths with the aim to recreate virtual objects in a physical three-dimensional space. [Haines and Haines, 1992] first applied holography to the field of computer graphics. HMD AR displays have applications in a variety of scenarios and are currently used in some of the modern headsets, such as Microsoft Hololens or Meta 2. This thesis focuses on mobile AR, but the technology researched on it, could be applied to HMD AR devices in the future.

Spatial Augmented Reality

Spatial Augmented Reality (SAR) alters real-world objects through the use of projectors. This technology allows modification of shapes and textures of physical objects without the use of a screens or display. This technique is also known as *Projection* or *Video Mapping*.

An early example of projection-based augmentation can still be found in *Disneyland California Park* with projected heads in the *Haunted Mansion*. These projections include *Madame Leota*, the ghostly head appearing inside a crystal ball in the séance scene, and the quintet of singing busts from the graveyard scene. [Mine et al., 2012] reviewed the use of SAR in Disney theme parks. Nonetheless, the term of SAR is much more recent and it was not first introduced until [Raskar et al., 1998b]. [Raskar et al., 1998a] also explored and foresaw how this technology could revolutionize future offices.

Projection mapping relies on the projector and target position to augment physical objects. To do so, the mapped image is masked according to the object's shape from a specific point of view. [Raskar et al., 2001] introduced this approach to enable animations on real-world objects through image-based illumination. To allow for interactive projections, [Beardsley et al., 2005] introduced hand-held projectors to enhance portability and provide a practical addition to mobile computing. SAR has applications in a variety of scenarios. [Ishii et al., 2002] used SAR for overlaying urban planning drawings. [Shelton and Hedley, 2002] used SAR to enhance geography teaching to undergraduate students. [Liao et al., 2010] guided MRI surgeries through the use of SAR projections. This thesis foresees the use of SAR for future use cases of IR.

Hand-held Augmented Reality

Technological improvements of the late 1990s and early 2000s allowed an extended use of AR and VR to a broader scope of researchers and users. Hand-held AR is a clear proof of such advancements, as it emerged from the first generation of mobile phones with cameras in

the early 2000s. At that time, [Wagner and Schmalstieg, 2003] described the first stand-alone AR system running in an unmodified hand-held device. Its main advantage over other types of AR is the fact that mobile devices are widely available to the general public.

A broad spectrum of potential applications have been introduced using hand-held AR. For instance, [Besbes et al., 2012] proposed this technology as a tool for industrial maintenance training. [Lehtonen et al., 2016] introduced an extended panorama tracking capable of performing context-aware applications. [Magnenat et al., 2015] used hand-held AR as a tool to enhance creativity and child development. This thesis mainly focuses on using hand-held AR to enable IR as a framework for distributed collaboration.

2.5.2 Spanning the Reality-Virtuality Continuum

Creating experiences that interact across the entire spectrum of the RVC is a relatively isolated field of research. [Davis et al., 2003] presented the first research in this domain. They proposed a continuum of virtual environment experiences for the book *Alice's Adventures in Wonderland*. They presented an adapted version of this classic which transitioned between realities throughout the story. This initial prototype was followed by [Nijholt, 2005] with work on virtual avatars along the RVC. Participants were captured in the real-world and ported to the virtual one allowing the participants to experience themselves throughout the continuum. [Andre, 2006] further extended this by engaging conversations between synthetic and human agents.

Our research presents a novel approach to this concept in which the user can interact throughout the RVC using only a mobile device in a single scenario (see section 5.2). Previous experiences required multiple sets and devices to be able to experience the different realities across the continuum, which does not allow a seamless transition between them.

2.5.3 Gaming between reality and virtuality

Gaming between reality and virtuality is a form of playing that combines components from the real and virtual world. This form of play is often found in location-based games. [Kasapakis et al., 2013] evaluated the design aspects founding them to beneficial for a broad scope of uses. [Jantke et al., 2013] applied them to *Aliens on the Bus*, a family pervasive game in AR. [Zarzycki, 2012] proposed *Urban Games*, with the aim to simultaneously inhabit real and virtual cities. Further, [Karl et al., 2010] proposed *Undercurrents*, a computer-based game-play tool to support tabletop role-playing. Popular examples of *Trans-Reality* games are *Pokemon GO* by [Niantic and Company, 2016] or *Birdly* by [Rheiner, 2014]. We extend this concept with what we define as *Multi-Reality games*, where the player evolves across the entire RVC (see section 5.2).

2.5.4 Stop Motion Animation

Stop motion animation techniques can be a demanding process for movie productions, which can lead up to many person years of effort. An approach to solve this was introduced by [Zheng and Wang, 2013] through a hand-driven video-based interface. Their method created automatic temporal interpolation between key-frames in a two-phase based capturing and processing work-flow. It achieved smooth transitions between poses, but did not contemplate reconstructing over occluded areas when those were not present on the initial frame of reference.

[Emerson, 2015] proposed an approach based on an additive manufacturing pipeline. They produced puppets and faces *en masse* to reduce effort and costs. [Zilly et al., 2016] introduced a light-field method to decrease costs using a static multi-camera system. It provided an approach for speeding up post-processing effects and enabled changes in the depth-of-field and smooth camera moves. Nonetheless, this approach did not address the

need to manually create in-between deformation frames and requires a calibrated camera array.

This thesis introduces IR as a tool that reduces the number of 3D printed puppet components required in a traditional stop motion animation (see section 5.3). To do so, the in-between frames of the key poses are computer-generated, having the only requirement to have the key poses physically present. To achieve this transition, we use AR to perform deformation of the surface samples projected from the video feed. Further, a diminished reality technique is implemented to deal with virtual deformations which digitally reveal visually erroneous surfaces obscured by the physical prop.

2.6 Summary

IR relates to a variety of research areas in AR to integrate a CMR system as defined by [Billinghurst and Kato, 1999]. This thesis draws from [Zünd et al., 2015], [Fairchild et al., 2017] and [Steed et al., 2012] as related examples of MR cooperation to enable collaboration through the remote possession of toys that come to life. [Zünd et al., 2015] introduced a system in which multiple collaborators could simultaneously enhance creativity in AR. [Fairchild et al., 2017] and [Steed et al., 2012] proposed to embody processes of capturing remote participants beamed into the shared space with visual depictions of them displayed digitally in mixed reality. Our system provides a real-world physical intermediary for the remote person's presence with animated expression synchronized with transmitted vocal audio.

The use of physical objects as avatars for remote collaboration has been conceived previously. [Sekiguchi et al., 2001] introduced a RUI for interpersonal communication using robots as physical avatars. Drawing from the research introduced by [Karsch et al., 2011], in which they achieved to render photo-realistic synthetic objects into legacy images, we aim to create the illusion of movement on static puppets by seamlessly rendering synthetic

objects into a real-world scene. To create such appearance, we draw from image retargeting techniques. This approach aims to retarget an image to a form with visual constraints meanwhile preserving its main features. In this thesis, we apply this concept to AR, in the case where we have known segmented 3D geometry of physical objects in the image and the ability to register the mesh pose in real-time using marker-less 3D object tracking by [Comport et al., 2006]. Drawing from related schemes, such as *Altered Reality* by [Leão et al., 2011], or *ClayVision* by [Takeuchi and Perlin, 2012], we introduce deformable object retargeting and shadow retargeting.

With the aim to create a seamless experience, we introduce a light-estimation approach that enables light-source detection using flat Fresnel lenses embedded in every-day objects. Drawing from the reflective chrome sphere method introduced by [Basri and Jacobs, 2001], we estimate light-source positions through a 180 degree planar Fresnel film. We embed this reflector in detectable objects without the need of additional gadgets.

IR is designed to account for a diversity of MR applications in which a collaborative environment can be highly beneficial for distributed participants. Following the work on computational imaging for stop motion animated video productions by [Zilly et al., 2016], this thesis introduces IR as a tool that reduces the number of 3D printed puppet components required in a traditional stop motion animation. To do so, the in-between frames of the key poses are computer-generated, having the only requirement to have the key poses physically present. Further, we present Multi-Reality Games, a novel game experience in which the user can interact throughout the RVC using only a mobile device in a single scenario. Drawing from previous related schemes, such an adapted version of the book *Alice's Adventures in Wonderland* by [Davis et al., 2003] that transitioned between realities throughout the story, or from avatars meetings in the virtuality continuum by [Nijholt, 2005], participants are able to transition across the RVC seamlessly in a unique device.

Chapter 3

Appearance Retargeting

This chapter addresses the gap between virtual and physical worlds using interactive AR to convincingly re-animate physical objects through digital displays. Section 3.1 introduces this novel approach for interactive AR. Section 3.2 introduces the *Props Alive Framework*. This framework draws from *Image Retargeting* techniques to animate physical puppets. Section 3.3 details how to seamlessly estimate scene light sources to correctly retarget animated objects. Section 3.4 details how to achieve an illusion of movement from the real-world object through image retargeting techniques using AR. Section 3.5 shows how already present shadows from static real objects in the scene can be retargeted according to a virtual overlaid AR movement over time. Section 3.6 provides a discussion of the work presented in the chapter. Section 3.7 concludes with a summary of the chapter.

3.1 Introduction

Producing visually realistic compositions at interactive rates in AR is becoming of increasing importance in both research and commercial applications. It requires relighting techniques to achieve plausible integrations between virtual objects and real-world scenes. Such reality mixing, when done in mobile computing, becomes specifically challenging due to the limited

amount of available resources. Nonetheless, this seamless integration between virtuality and reality is becoming more achievable due to the recent advances in high-performance mobile computing. Such cutting-edge algorithms achieve enhanced real-time capabilities for MR applications in commercial mobile devices. This trend enables our goal to combine the virtual and physical worlds seamlessly together and create a convincing illusion of movement for normally static and inanimate objects. This chapter introduces the following contributions:

- we introduce a software framework, titled *Props Alive*, that provides the structure required to bring objects to life in mobile AR (see section 3.2, [Casas et al., 2017]).
- we introduce a novel light-estimation approach employing a flat light probe fabricated using a flat reflective Fresnel lens. Unlike three-dimensional sampling probes, such as reflective spheres, Fresnel lens can be incorporated into any type of object, such as a book or product packaging, providing a seamless light-estimation approach for AR experiences (see section 3.3, [Casas et al., 2019]).
- we describe *Deformable Object Retargeting*. This method achieves an illusion of movement from the real-world object through image retargeting techniques using Augmented Reality (see section 3.4, [Casas et al., 2017]).
- we introduce appearance reconstruction, a method that is capable of reconstructing areas not present in the physical reference model when revealed in AR (see section 3.4.2, [Casas and Mitchell, 2019]).
- *Shadow Retargeting* is presented. This method is an efficient and focused approach in which already present shadows from static real objects in the scene are retargeted according to virtual overlaid AR movement (see section 3.5, [Casas et al., 2018b]).

3.2 Props Alive Software Framework

Bringing props to life in a realistic manner using AR is a complex task. In order to tackle this, we have created *Props Alive Software Framework*, an ad-hoc framework developed exclusively for such propose. We base our system design on the framework for information visualisation introduced by [Kennedy et al., 1996], which generically considers user interaction with data. We implement our framework with [Unity, 2019]. This technology provides a cross-platform solution that enables development for an extended variety of devices with rapid prototyping. In this particular case, our framework runs on contemporary mobile devices as a deployable app. In addition, [Vuforia, 2019] is used to support tracking of the real-world object in the scene and delivers the video image and pose registration parameters to our *Props Alive framework* components. As figure 3.1 indicates, our framework is split into three main core parts:

- Flat Fresnel Lens Light-Source Estimation (see section 3.3).
- *Object Retargeting* (see section 3.4).
- *Shadow Retargeting* (see section 3.5).

3.3 Flat Fresnel Lens Light-Source Estimation

To accurately account for pose deformations in the scene, we must estimate the main light source directions that cast shadows to the object that we intend to retarget. In this research, we introduce a method that allows automatic detection of distant light sources using a flat grooves-out reflective Fresnel lens. These use a concentric ring arrangement to form an optical equivalent of a solid 3D light probe as the ones introduced by [Debevec, 1998].

The proposed solution covers a more flexible and adaptable method for light estimation than previous approaches as very little configuration is needed. *The Shading Probe* by [Calian

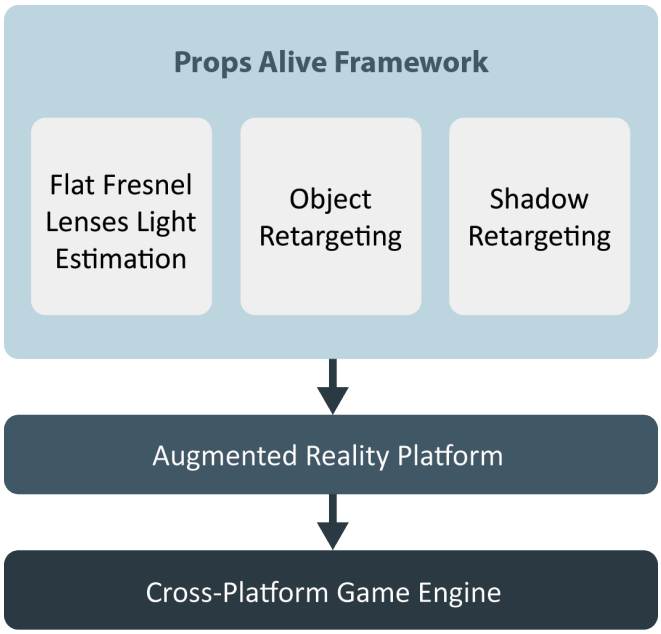


Fig. 3.1 The data flow diagram details how our *Props Alive framework* is built on top of an augmented reality platform ([Vuforia, 2019]) and a cross-platform game engine ([Unity, 2019]).

et al., 2013], required a 3D-printed and pre calibrated light probe. [Castro et al., 2012] and [Nowrouzezahrai et al., 2011] relied on using a reflective light-probe sphere to pre-calibrate the scene. [Knorr and Kurz, 2014] and [Calian et al., 2018] used human faces as light probes for coherent rendering of virtual content. However, the accuracy of the face tracking is a limiting factor on the quality of the recovered illumination. Moreover, the method is limited to applications where only frontal face views are available. A variant of light estimation from faces further developed in [ARKit, 2019].

[Mandl et al., 2017] used *Convolutional Neural Networks (CNN)* to relight every-day objects. However, this technique required to previously train a CNN with finite number of light configurations in a given 3D object. For each view, spherical harmonics were extracted and stored in a lookup table. Here, instead, we propose a method in which the flat Fresnel lens can be integrated into any time of objects with increased design freedom and flexibility than previous approaches. Section 3.3.1 describes the method that addresses anisotropic

artefacts for Fresnel lens reflection. Section 3.3.2 details the light estimation algorithm implementation.

3.3.1 Anisotropic Reflection Handling

Our method for light estimation relies on flat grooves-out Fresnel lens embedded on pre-registered objects to detect distant light sources in the scene. We use marker-less tracking to recognise objects that have embedded lenses. These detection targets are mesh, which enables us to segment the object's region in which the flat Fresnel lens is embedded. It draws from the light source estimation algorithm using reflective chrome spheres from [Basri and Jacobs, 2001], to extend its reproducibility using flat Fresnel lens films.

In essence, a flat grooves-out Fresnel lens film is the reproduction of a chrome-ball sphere in a two-dimensional surface. Such film is able to reproduce the reflection of a 180-degree sphere over a plane. Light beams reflecting over this lens are refracted by the Fresnel coefficients of the surface film. These describe the ratio of the reflected and transmitted electric fields to that of the incident beam. As such complex coefficients shift the phase of the pulse between waves, highly saturated anisotropic reflections appear. In order to segment the light-source origin within the Fresnel lens, and avoid false positives given by anisotropic reflections, we additionally capture the Fresnel lens with a very low exposure setting. This procedure reveals the light-source origin due to its highly saturated values. Section 3.3.2 details the algorithm's steps to estimate the light source in the scene using flat Fresnel lenses.

3.3.2 Light Penumbra and Direction Estimation Algorithm

Using [Vuforia, 2019]'s marker-less tracking, we can make virtual-world space coordinates coincident with the real-world detected object in the environment. As we use mesh-prior, meaning that we place the virtual mesh on the exact same location as the physical object, we pre-establish the geometry that contains the flat Fresnel lens film for segmentation. This

allows us to detect the image region that we would use for reflectance estimation. Once the region where the flat grooves-out Fresnel lens is segmented, we first determine its center by calculating its radius r. We do so by taking the maximum distance (equation 3.1) from segments pixels ($p[n]_{(x,y)}$) to the flat Fresnel lens center ($Fc_{(0,0)}$) within region n. We account for image perspective correction by obtaining the virtual AR camera field of view, which is coincident with the real camera from the mobile device. This radius r is used to compute the *penumbra* regions of the virtual shadow when an area light source is present in the scene.

$$r = \max \left\{ \sum_{n=1}^{p} dist(Fc_{(0,0)}, p[n]_{(x,y)}) \right\} \tag{3.1}$$

Within this area, we look for specific pixels where RGB channels are highly saturated following equation 3.2. $\vec{h}$ is a two-dimensional vector that contains the co-ordinates of highly saturated pixels. p is the total number of pixels within the segmented region. $\hat{\mathbf{R}}$, $\hat{\mathbf{G}}$, and $\hat{\mathbf{B}}$ are the normalized channels of each pixel p within region n. t is the predefined saturated threshold for detection of highlight regions.

$$\vec{h_{(x,y)}} = \sum_{n=1}^{p} p[n]_{\hat{\mathbf{R}}} > t \wedge p[n]_{\hat{\mathbf{G}}} > t \wedge p[n]_{\hat{\mathbf{B}}} > t \tag{3.2}$$

The Fresnel lens follows the same principle of geometrical optics used for other commercial lenses. The focal length and aperture are the two main parameters of this principle, which define the center of the reflection based on the prism angle α. We estimate this center ($c_{(x,y)}$) by calculating the average co-ordinates of adjacent saturated pixels $p[n]_{(x,y)}$ within a region. We do this for each highlighted area on the image following equation 3.3. This approach is capable of detecting the N main light sources of the scene as we are able to estimate its center $c_{(x,y)}$ for each real-world light source and then apply them to *Shadow Retargeting* (see section 3.5).

$$c_{(x,y)} = \frac{\sum_{n=1}^{|h|} p[n]_{(x,y)}}{|h|} \tag{3.3}$$

Once the center of the highlight is defined, we use the normal vector of the flat Fresnel lens surface to project $c_{(x,y)}$ along its z axis. We then use the radius r to compute a hemisphere from the flat Fresnel lens center ($Fc_{(0,0)}$). The intersection point between both surfaces in the euclidean space is noted as $P_{(x,y,z)}$. Since the normal N of the point $P_{(x,y,z)}$ over the surface of the hemisphere is known at any given point, we can reconstruct light sources directions in the environment for each projected highlight center. To obtain the light-source direction $L_{(x,y,z)}$, we follow equation 3.4, in which N is the normal vector at the highlight center point, and R is the reflection vector, $R = (0,0,-1)$.

$$L_{(x,y,z)} = 2(N \cdot R)N - R \tag{3.4}$$

3.3.3 Results of Light Source Estimation

Figure 3.2 shows the results of estimated light-source directions following the approach described in section 3.3.2. *Row 1* displays the environment lighting of the scene captured through flat Fresnel lenses. *Row 2* shows underexposed images capturing the light-source point represented as highly saturated pixels within the image. *Row 3* displays a visualization of the light-estimation algorithm following the projection from a flat Fresnel lens to a virtual superimposed hemisphere. *Row 4* shows ground-truth shadows under real-world lighting. *Row 5* shows synthetic hard shadows cast from the centre of the highly saturated region with the flat Fresnel lens using our light-estimation algorithm. *Row 6* shows a comparison between ground truth and synthetic hard shadows. *Row 7* uses the radius of the highly saturated region to compute the penumbra area using PCSS. As can be seen in the *row 8*, our algorithm is able to consistently estimate the light direction of the scene. Inaccuracies are prone to be present on the penumbra area. This can be observed in the normalized comparison between

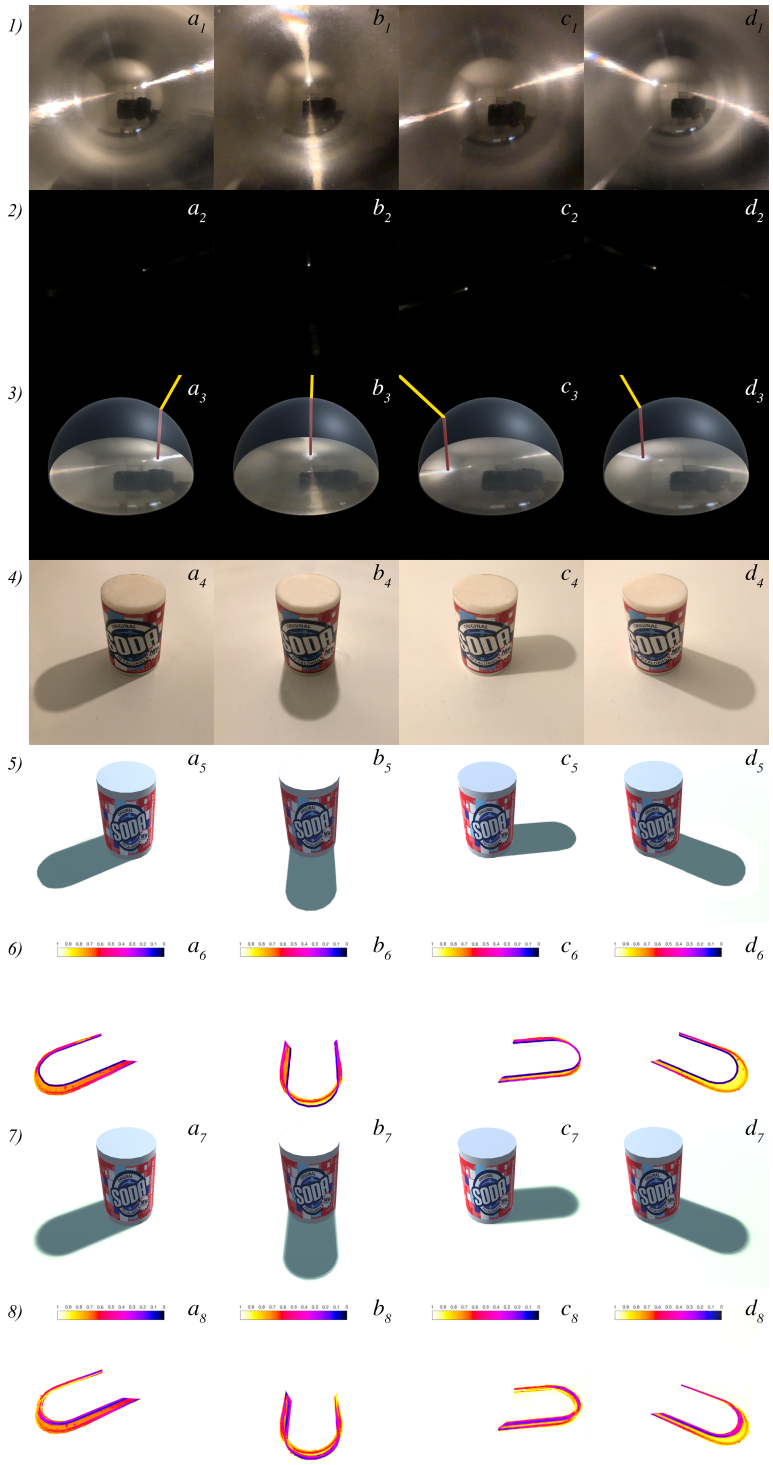


Fig. 3.2 **(1)** Environment lighting captured through flat Fresnel lenses. **(2)** Underexposed images that capture the light-source point represented as highly saturated pixels within the image. **(3)** Visualization of light-estimation algorithm. **(4)** Ground-truth shadows. **(5)** Synthetic hard shadows cast from the centre of the highly saturated region with the flat Fresnel lens. **(6)** Comparison between ground truth and synthetic hard shadows. **(7)** Synthetic soft shadows using the radius of the highly saturated region for the penumbra area using PCSS. **(8)** Comparison between ground truth and synthetic soft shadows with reduced error.

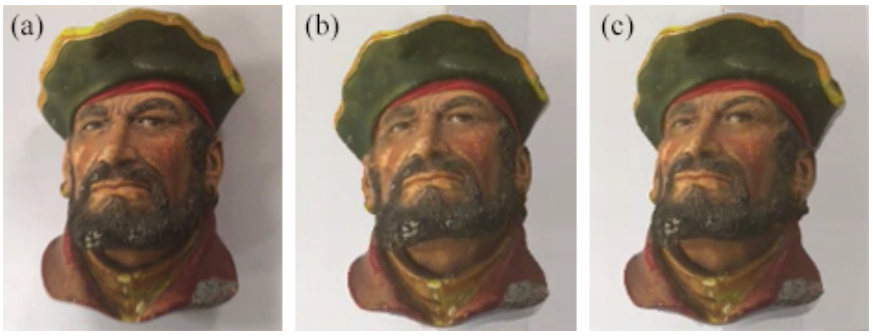


Fig. 3.3 **(a)** Hand-painted pirate head *Bossons* plaster model. **(b)** The virtual pirate head rendered at the same pose. **(c)** The virtualretargeted pose using *Deformable Object Retargeting*.

ground truth and those that are synthetically generated. Such inaccuracy is low and could further be reduced by enhancing the computation of the virtual light-source radius.

3.4 Deformable Object Retargeting

Our framework requires to have a physical puppet, either 3D printed or manually formed, with its key physical poses present in the real-world. In the case of our pirate head, (see figure 3.3), a hand-painted *Bossons* plaster model. Our digital equivalent meshes are created through photogrammetric software. To animate our real-world puppet, we rig and retopologize the mesh obtained from the photogrammetric scan using a standard 3D editor software. We use image markers or point cloud registration to detect the real-world object location. Virtual meshes use that tracked position to match the same location in the virtual coordinate system and use the resulting texture coordinates for warping (see section 3.4.1) and reconstructing (see section 3.4.2) the live camera video feed texture.

3.4.1 Texture Warping

In order to perform real-time deformed texturing, we place a proxy object locator in the pivot position of the real-world marker and we assume that no *root* movement will be performed on it. To achieve an accurate registration of the physical object, we use marker-less tracking

for augmented reality using [Vuforia, 2019]. This consists of a system based on point clouds that recognises rigid, opaque physical objects. When detected, this point cloud registers the origin of the world-space co-ordinate system. The virtual mesh is placed coincident at the origin of this Euclidean space. Using this procedure, we can simulate where the physical object is in our virtual co-ordinate system. We can estimate the position of the ground plane also from the origin of the Euclidean space, since we assume that the detected object is in contact with the ground. Texture coordinates of the overlaid animation are back projected to the initial rest pose. This achieves real-time deformed texturing sampled from the real world object's vertex locations in the projected video camera source image. With the aim to obtain coherency on the deformation of the texture, we use a direct map between initial rest pose vertices and animated ones. Look-up coordinates are updated on a per frame basis, in accordance to the position of the camera and the virtual mesh animation.

Our approach enables an optimized lookup of the animated vertex position as gives us a $\mathcal{O}(1)$ constant access time. Any changes applied to the surroundings of the physical object will perform its expected lighting effects to the virtual mesh. This method holds plausibly for small-proximity displacements as seen in figure 3.5. Large translations in the virtual object from the physical reference position cause lighting inconsistencies with spatial variations in the environment.

3.4.2 Texture Reconstruction

For areas not present in the physical reference model, we need to reconstruct the puppet appearance in a plausible way. This is often the case for teeth and the interior part of the mouth which only gets revealed when the puppet retargets certain facial expressions (i.e. mouth opened). As our method uses a three-dimensional mesh that matching the physical object shape, we can accurately establish beforehand regions that will be revealed and are

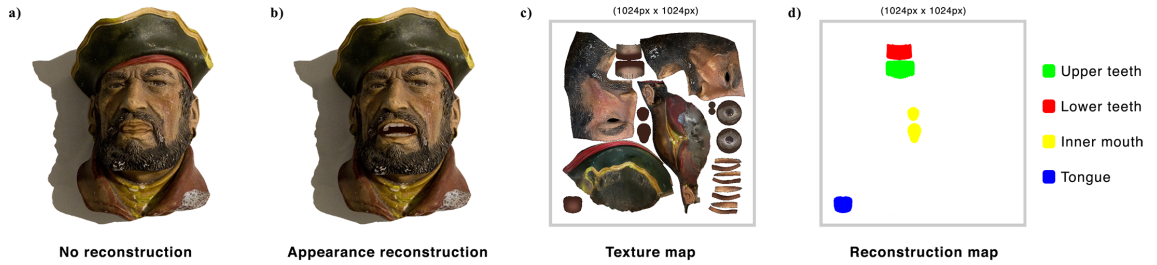


Fig. 3.4 **(a)** Frame of an animated physical puppet in AR without reconstructing occluded areas that become revealed. Teeth and interior parts of the mouth get rendered with the aspect of lips. **(b)** With appearance reconstruction, these parts get rendered plausibly from alike regions of reference. **(c)** Unwrapped texture map of the virtual mesh. **(d)** Color encoded texture map segmenting regions to be reconstructed in real-time.

not visually seen in the reference puppet. Those areas need to be paired off-line with an alike albedo estimation in order to get reconstructed plausibly in real-time.

To do so, we unwrap the texture map (see figure 3.4c) of the model onto the mesh. Through the visualization of the mesh in a standard 3D editor, we identify and segment occluded regions of the mesh that have no correspondence on the real-world reference and will need to be reconstructed when blend shapes values are applied to the augmented toy. We do so by pre-identifying vertices from the geometry that will need to be reconstructed, and hence, *inpainted*. We segment by area according to the element that they represent. Each element is color encoded in a texture map (see figure 3.4b) and paired to an area present in the real-world object that contains a desired similar appearance. This method applies only for cases in which the desired albedo can be sampled from a region that contains a similar appearance. Nonetheless, this in essence generalises to most puppets and humanoids as these tend to mimic and resemble the outlook and characteristics of humans.

3.5 Shadow Retargeting

Shadow retargeting leverages the constraints of shadow visibility and appearance with known geometry to efficiently steer source shadow samples for retargeted reconstruction with high

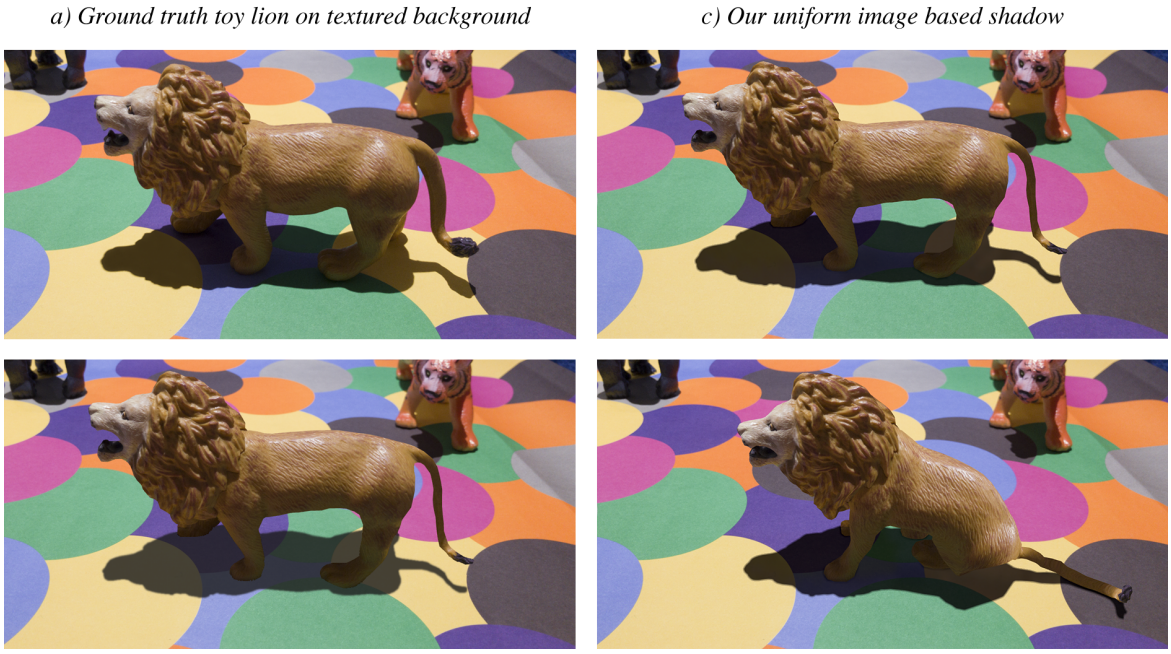


Fig. 3.5 (**a**) Reference physical toy lion as observed by a mobile phone camera. (**b**) The virtual lion rendered at the same pose with (**b**) a regular shadow map approach and (**c**) with *Shadow Retargeting*. (**d**) Its movement retargeted according to a sequence of geometry poses in AR towards seamless appearance preservation.

quality. Section 3.5.1 offers an overview of the method. Section 3.5.2 details the warping of the physical projected shadow. Section 3.5.3 illustrates our sampling search for shadow reconstruction. Section 3.5.4 describes the retargeting model. Section 3.5.5 gives an overview of the *Shadow Retargeting* algorithm. Section 3.5.6 contains our procedure for retargeting soft shadows. Section 3.5.7 details the use of auxiliary masks for optimized rendering performance. Section 3.5.8 performs depth coherent selection for multiple candidates eligible to be rendered in the same fragment. Section 3.5.9 handles multiple shadow overlapping. Section 3.5.10 deals with non-grounded occluders. Section 3.5.11 handles multiple receiver surfaces. Section 3.5.12 shows how our approach deals with background inpainting. Finally, section 3.5.13 showcase *Shadow Retargeting* results.

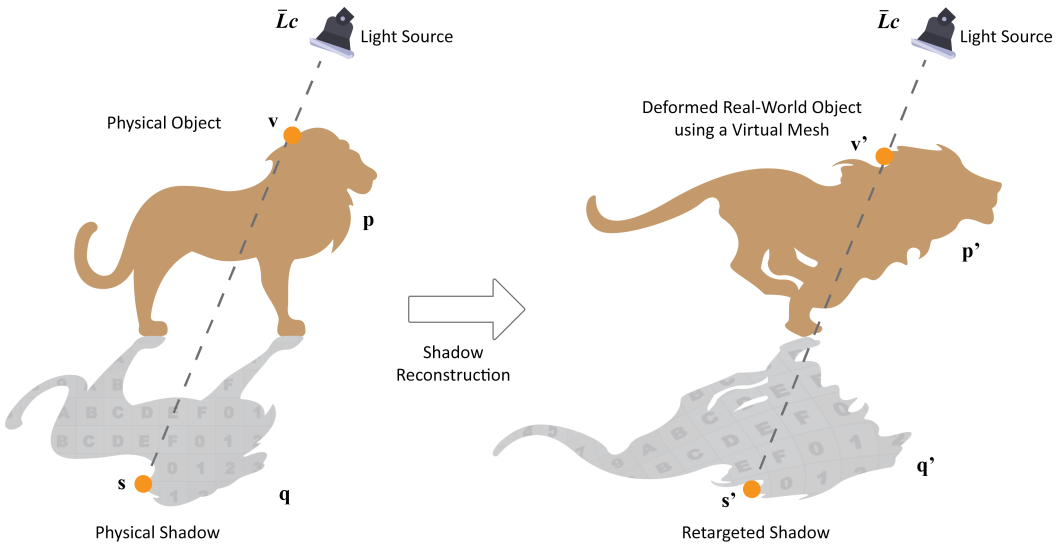


Fig. 3.6 **(p)** is the real-world static object. **(q)** is the physical projected shadow. **(v)** is a vertex from the coincident overlaid mesh. **(s)** is *v* being projected to the floor. **(p')** is the deformed real-world object. **(q')** is the retargeted projected shadow. **(v')** is same initial vertex *v* being animated over time. **(s')** is *v'* being projected to the floor.

3.5.1 Virtual Shadow Re-sampling Overview

Our approach synthesizes virtual shadows directly by sampling the image that contains the shadow of the object presented in the real world (see figure 3.5c). As the virtual mesh is animated, we warp the real shadow to retarget its appearance to the virtual one (see figure 3.5d). This process approximates the shadow as being directly linked to the projection along the principal light direction of the caster on the receiver. Our method requires to estimate the direction of the *n* main light sources in the scene. We compute this information at the time of initialization following the method described in section 3.3. Our method holds plausibly for small proximity displacements and exploits visibility as a smooth function of the given geometry displacement for natural lighting warped reconstruction.

3.5.2 Shadow Warping

Our goal is to warp the shadow from the image to its displaced version. We rely on prior knowledge of mesh geometry and register it with its 3D printed or scanned physical version.

Fig. 3.7 (**a**) Mask for physical-object pose. (**b**) Mask for physical-shadow pose. (**c**) Mask containing the valid shadow area eligible for appearance sampling.

In a first step, we project the object's mesh vertices in their world-space positions along the light direction on the receiver geometry. We then re-project to image space in order to associate each mesh vertex to its position on the shadow. As a second step, we project the animated mesh to the ground using the position of those vertices in the virtual deformed pose and use previously mapped texture co-ordinates to interpolate the original shadow across the virtual shadow (see figure 3.6). This approach supports unique and multiple receiver surface (see section 3.5.11).

3.5.3 Shadow Reconstruction

The shadow warping step acquires as much as possible of the whole real shadow visible in the image. In the base case, it does not take into account any further shadow occlusion (see section 3.5.9 for overlapping shadow considerations). We are able detect occlusion from the mesh since it is registered in the image. We then search for another reference point to synthesise the reconstructed shadow appearance.

We synthesise a mask of eligible shadow area valid for sampling, by first rasterizing the projected shadow of the real-world object (see figure 3.7b), and then rasterizing the object itself to remove its own shadow occlusion (see figure 3.7a). This results in an eligible shadow area (see figure 3.7c).

Where the shadow warp would sample an invalid texel, we perform a discretized concentric ring search to find the closest valid (i.e., onocludded-shadow) texel (see figure 3.8).

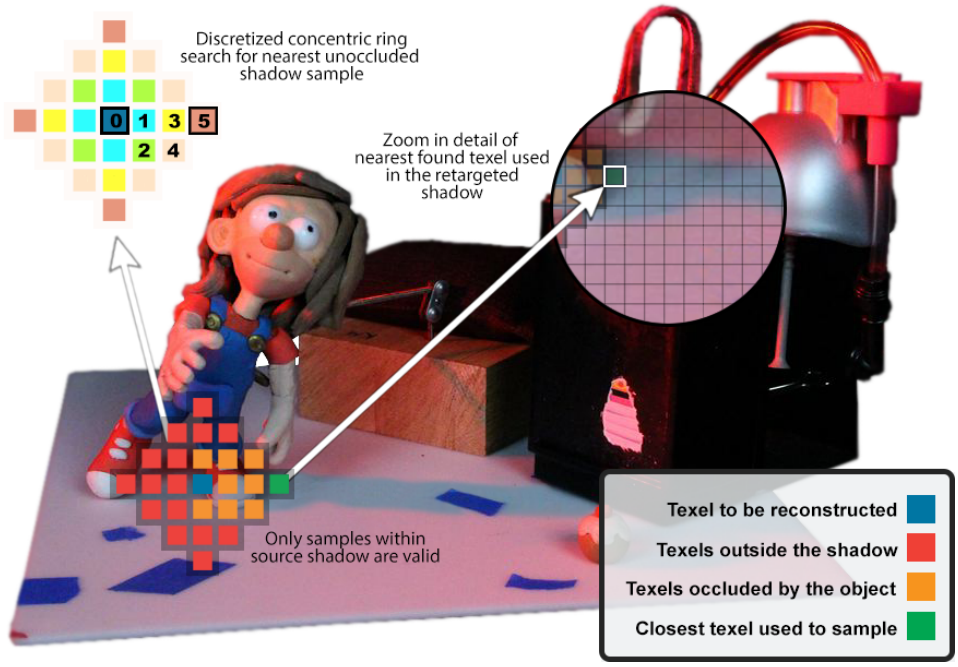


Fig. 3.8 Discretised concentric-ring search algorithm for plausible shadow reconstruction. Red samples are trivially masked and rejected as ineligible. The closest eligible texel, illustrated as a green sample, was used for appearance sampling.

The search samples all the candidates for each iteration at a time. It stops once it finds a candidate that is valid. In the case of having multiple eligible candidates, the one closest to the invalid texel is sampled. To compensate for appearance discontinuities resulting from the inpainting process, we performed a smoothing pass using an edge-aware box linear filter. Samples outside the shadow are discarded in this pass, as we assume a uniform light source.

3.5.4 Retargeting Model

By retargeting shadows, the albedo of the receiver can change over non-regular backgrounds. To take into account materials with non-uniform albedo, we need to relate the outgoing radiance of the receiver in shadow with its radiance in light. The calculated ratio is applied to a new point in the retargeted shadow. For both reference points, in shadow and light, we assume that there is no emissive radiance or subsurface-scattering events. We express the outward radiance L_o as a derivation from the rendering equation 3.5 introduced by

[Kajiya, 1986]. Our method derives a ratio formulation between shadow and non-shadow observations.

$$L_o = \int_\Omega f_r(\mathbf{x}, \omega_i, \omega_o) L_i(\mathbf{x}, \omega_i) (\omega_i \cdot \mathbf{n}) \, d\omega_i \tag{3.5}$$

$f_r(\mathbf{x}, \omega_i, \omega_o)$ is the *Bidirectional Reflectance Distribution Function (BRDF)*, the proportion of light reflected from ω_i to ω_o at position $\mathbf{x}$. $L_i(\mathbf{x}, \omega_i)$ is the radiance coming toward $\mathbf{x}$ from direction ω_i. $\omega_i \cdot \mathbf{n}$ is the incidence angle factor. ω_i is the negative direction of the incoming light. Assuming a diffuse BRDF (f_r), the incoming light is scattered uniformly in all outgoing directions. Therefore, in this case, the BRDF does not depend on the incoming ω_i and outgoing ω_o light directions, and becomes a constant determined as ρ (see equation 3.6).

$$f_r(\omega_i \rightarrow \omega_o) = \rho \tag{3.6}$$

Given our diffuse reflectance ρ_d, we are constrained to a constant subset of full BRDFs. The diffuse BRDF and diffuse reflectance are related by a π factor (see equation 3.7).

$$\rho_d = \pi\rho \tag{3.7}$$

Since we can not evaluate the exact irradiance of the real-world scene without accurate light estimation, we assume a constant uniform light source in a position ($\mathbf{x}$) between reference points in shadow (E_{i0}) and light (E_{i1}) (see equation 3.8).

$$\begin{aligned} E_{i0} &= L_{i0}(\mathbf{x}, \omega_i) \\ E_{i1} &= L_{i1}(\mathbf{x}, \omega_i) \\ E_{i0} &= E_{i1} \end{aligned} \tag{3.8}$$

Therefore, under these conditions, the ratio of outgoing radiance is conserved for any uniform illumination in the same position (**x**) in light (L_{o0L}) and shadow (L_{o1S}). This ratio is simply computed as the diffuse BRDF in light (ρ_{d0}) divided by the diffuse BRDF in shadow (ρ_{d1}) (see equation 3.9).

$$L_{o0L}/L_{o1S} = \rho_{d0}/\rho_{d1} \quad \forall E_i \tag{3.9}$$

We use this ratio to compute the new point in shadow (L_{o0S}) drawing from the same reference point in light (L_{o1L}). Both reference points are photometrically calibrated and gamma corrected (see equation 3.10).

$$L_{o0S} = \rho_{d0}E_{i0} = \frac{\rho_{d0}}{\rho_{d1}}\rho_{d1}E_{i0} = \frac{L_{o0L}}{L_{o1S}}\rho_{d1}E_{i1} = \frac{L_{o0L}}{L_{o1S}}L_{o1L} \tag{3.10}$$

3.5.5 Shadow Retargeting Algorithm

Our method retargets shadow texels given a corresponding deformation of the geometry for each light source in the scene (see algorithm 1). When the shadow is projected onto a textured background, we reconstruct the shadow texel if occluded (see section 3.5.3), and apply the retargeting model (see section 3.5.4). To do so, the outgoing-radiance ratio is calculated by comparing the same texel in both light and shadow. Subsequently, this proportion is taken into account for the texel in the retargeted shadow. In the event that the retargeted shadow reveals areas that were initially in the shadow, we directly process the previously obtained texel from the observed background image.

In the case of a uniform background on the receiver, we simply warp the shadow if the texel is inside the *umbra* area from the appearance sampling mask (see figure 3.7c). Otherwise, we perform a discretized concentric ring search and reconstruct the shadow (see section 3.5.3).

Algorithm 1: Shadow-Retargeting Overview

```
foreach light source in scene do
    foreach shadow receiver in scene do
        if non grounded occluder then
            compute distance to receiver
            project shadow to receiver
        foreach projected target shadow texel in image do
            if receiver is textured background surface then
                apply retargeting model
                if soft shadow edge region then
                    return retargeted PCSS texel
                else
                    return retargeted shadow texel
            else if plain color uniform backdrop then
                if texel in umbra area then
                    return warped shadow texel
                else
                    if soft shadow edge region then
                        return reconstructed PCSS Texel
                    else
                        return reconstructed Shadow Texel
```

3.5.6 Retargeted Soft Shadows

Soft shadows are characterised by two main parts, *umbra* and *penumbra*. While the *umbra* represents the area in which all the rays emitted by the light source are occluded, the *penumbra* only represents those rays that are partially blocked. This is of particular importance when we aim to achieve photo-realism during the retargeting of a soft shadow as we are in need to reconstruct both, the *umbra* and *penumbra* areas (see figure 3.10). To do so, we use a multi-pass approach in which we first reconstruct the *umbra* and then recreate the *penumbra* using *Percentage Closer Soft Shadows (PCSS)* by [Fernando, 2005]. The kernel used to soften the shadow edge is determined by the distance to the first blocker of the light source using an mobile optimized edge aware filter introduced by [Chen et al., 2007]. To avoid losing texture details in the albedo, we apply a box linear filter to the shadow term L_{o1S} before applying albedo scaling term $\frac{\rho_{d0}}{\rho_{d1}} = \frac{L_{o0L}}{L_{o1L}}$.

3.5.7 Shadow Composition Masks

To ease rendering for our AR scenario, we make of use of auxiliary masks rendered as textures. These masks, as introduced earlier in section 3.5.3, are used to detect the position of a physical object in the scene. When creating these auxiliary masks, we create an instance of the mesh in its initial position and render it to a specific channel bit-mask component. Figure 3.7a shows the auxiliary masks in which the position of the physical object is shown. The same procedure is used to segment the physical projected shadow. In this case, we make use of the implementation described in section 3.5.5 returning only binary values, as shown in figure 3.7b.

3.5.8 Depth Coherent Selection

Our method warps the original shadow of the image to the virtual displaced one. When deforming, several candidates on the warped shadow may be eligible to be rendered in the

same projected fragment. If that happens, our method performs depth ordering according to the closest occluder. Therefore, after rasterizing the object, the closest one is represented in the depth buffer. This method preserves the shadow appearance, rendering the closest reconstructed sample when multiple candidates can be represented in the same fragment.

3.5.9 Handling Multiple-Shadow Overlap

When a multiple-shadow scenario occurs, each individual shadow has a unique mask in its initial and retargeted position. This allows us to detect shadow regions that overlap in their initial and deformed states. This detection is done at texel level and is evaluated when performing a discretized search for shadow reconstruction. Therefore, when there is an overlapping region, the inpainting sample is not sampled until all conditions in the search algorithm are fulfilled. For instance, if an overlapping region occurs in the retargeted binary masks, the search algorithm does not provide a sample until it finds a region where those initial masks also overlap. This approach preserves appearance in cases of multiple overlapping retargeted shadows (see figure 3.10).

3.5.10 Non-Grounded Occluder

In the case of the occluder not being in contact with the surface of the receiver, we compute the distance between both elements to accurately project the retargeted shadow. To do so, we make use of a ground-plane detector from [ARKit, 2019]. This mixed-reality library uses *Simultaneous Localization and Mapping (SLAM)* techniques to sense the world around the device. Using an *ORB-SLAM* based algorithm similar to [Mur-Artal et al., 2015], this library detects environment features and converts them to 3D landmarks. When multiple features converge in the same planar level, a ground-plane anchor is initialized. This anchor is used in our system as the world coordinates of the surface acting as shadow receiver. The projection

of the retargeted shadow is performed following sections 3.5 and 3.5.5. In the case of a soft shadow being cast, we recreate the penumbra area following section 3.5.6.

3.5.11 Multiple Receiver Surfaces

When multiple-receiver surfaces occur, such as a partially projected shadow over a ground plane and a vertical wall, we segment the original and projected shadow in subregions. Each receiver has its own subregion of the shadow assigned to it. We compute the retargeting algorithm detailed in section 3.5.5 for each receiver. Following a similar approach as in section 3.5.10, we use the ground-plane detector from [ARKit, 2019] to sense the environment of the device. In this case, we use planar and vertical anchors to retarget over multiple-receiver surfaces.

3.5.12 Background Inpainting

When animating a real-world object through AR, the background is revealed when movement is performed on the virtual animated object. This is known as the inpainting problem, in which a specific area of an image is desired to be patched according to its surrounding with a plausible outcome.

Real-time scene reconstruction with structural and perceptual consistency has demonstrated to be computationally expensive and early examples from [Bertalmio et al., 2003] were limited to off-line techniques. *PixMix* by [Herling and Broll, 2012] achieved inpainting in real-time with few background constraints and without the need of a multi-view approach. This algorithm iteratively minimized the spatial and appearance cost into a function achieving consistency over patterned backgrounds. Nevertheless, on-line performance was still limited to high end processors on portable computers.

The use of machine learning to solve the inpainting problem has been introduced recently. [Yeh et al., 2017] proposed a method to perform semantic image inpainting with deep

Fig. 3.9 (**a**) Observed background without the reference physical toy lion and shadow on it. (**b**) Real-world scene as observed by a mobile-phone camera. (**c**) Virtual lion in a retargeted pose and shadow with the revealed background reconstructed.

generative models. [Yang et al., 2017] achieved high-resolution image inpainting using multi-scale neural patch synthesis. Nonetheless, these techniques are still limited to off-line performance.

In the case of mobile devices, methods often require additional information. [Jarusirisawad and Saito, 2007] introduced an approach that used multiple hand-held devices to perform background inpainting. [Kawai et al., 2013] proposed a scene reconstruction approach based on depth segmentation for non-planar backgrounds. In our work, we use a background observation method, similar to *ClayVision* from [Takeuchi and Perlin, 2012], in which the user previously captures a representation of the background (see figure 3.9a). Masks 3.7a, 3.7b are sampled to the shader and evaluated in each fragment over a plane that entirely covers the region in which the physical object would render. Using these masks naively crops out the physical object and its projected shadow from the scene (see figure 3.9b). The segmented area is filled with the previously captured representation of the background in order to reconstruct the retargeted pose and shadow (see figure 3.9c). Once the image is reconstructed, we perform a linear interpolation between inpainted and unmodified pixels to achieve improved spatial and temporal consistency.

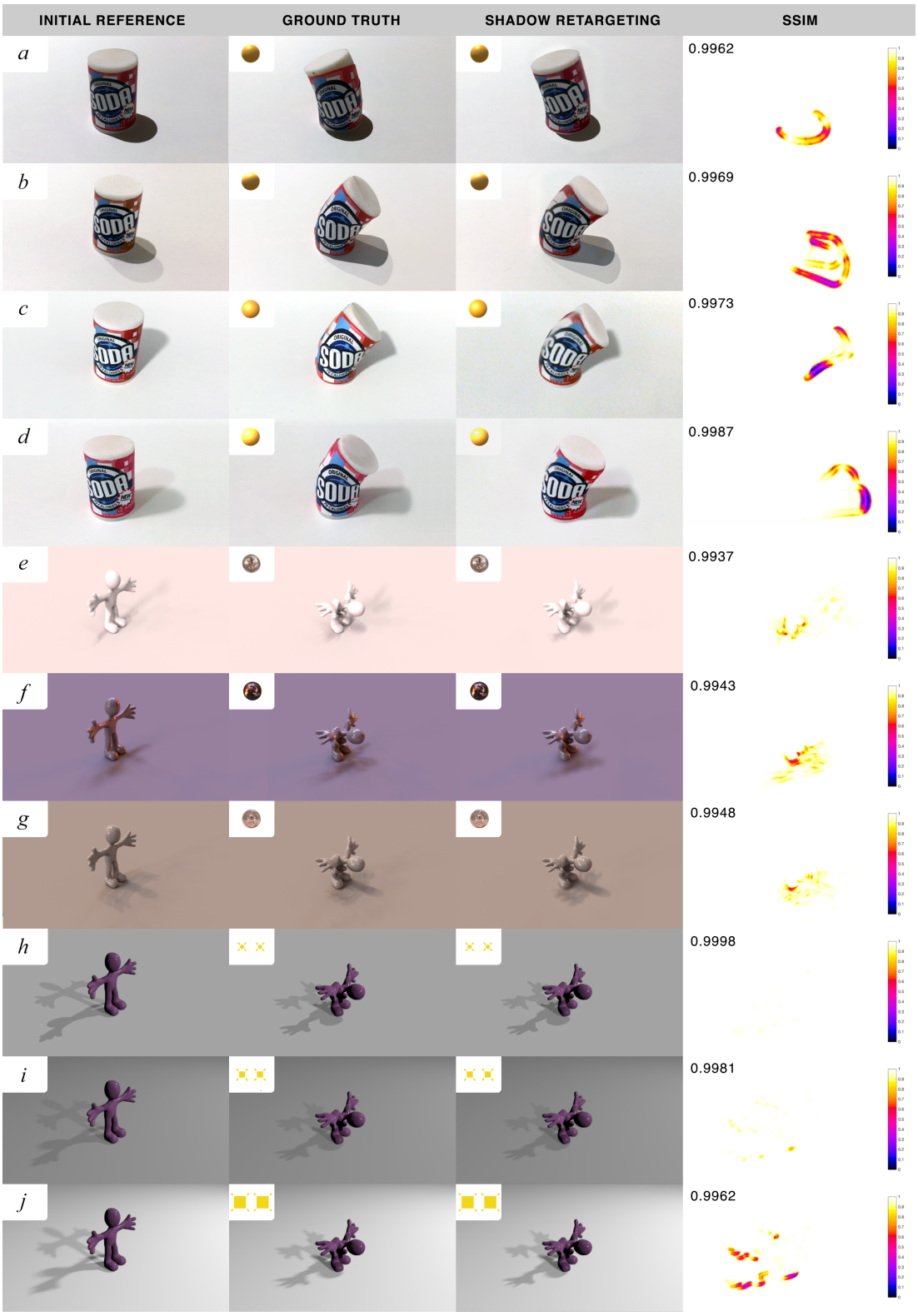


Fig. 3.10 Comparisons of ground truth with our shadow-retargeting scheme. (**a**–**d**) Real-3D printed nondeformed-object shadow retargeted with single/double light sources, and hard/soft shadows. (**e**–**g**) Synthetic renders with classic light probes showing retargeted complex shadows. (**h**–**j**) Effect of progression of point to area lighting retargeted from a rendered source image with character in bind pose.

3.5.13 Retargeted Results

Results of the shadow-retargeting approach are structured as follows. Section 3.5.13 shows comparisons between ground truth and our approaches. Section 3.5.13 details a time breakdown of a typical frame processed using our shadow-retargeting method. Section 3.5.13 shows visualisations of the inpainting technique for shadow reconstruction detailed in Section 3.5.3.

Ground-Truth Comparisons

Figure 3.11 shows a comparison between a ground-truth shadow generated by a physical occluder, the results using a simple estimated uniform shadow, and our method. As can be seen in the *Structural Similarity Index Metric (SSIM)* comparison, our method is closer to the ground truth since it is able to preserve the indirect lighting already present in the scene. Further, our ambient occlusion approximation further reduces error near contact points around the lion's feet. While the employed 3D-object-tracking registration and photogrammetric reconstructed model has accuracy that result in a visually stable animation in the video frame, slight misalignments result in the lines of higher error where discontinuities edges occur, e.g., in the lower portions of the zoomed in SSIM visualisations of figure 3.11.

Figure 3.10 shows a comparison between ground-truth shadows and our retargeting algorithm. *Results (a–d)* demonstrate the real-time capability of our algorithm for augmented reality. We performed texture deformation to animate the virtual-mesh and shadow samples to reconstruct occluded areas. Its key frames were 3D-printed in order to compare shadow retargeting with real-world shadows. SSIM results demonstrate good precision using our real shadow-retargeting technique, which may be improved with further accuracy of 3D marker-less object registration and tracking. *Results (e–g)* show results under High Dynamic Range (HDR) maps from [Debevec, 1998]. These demonstrate good precision in complex lighting scenarios, where we approximate multiple retargeted soft shadows. *Results (h–j)*

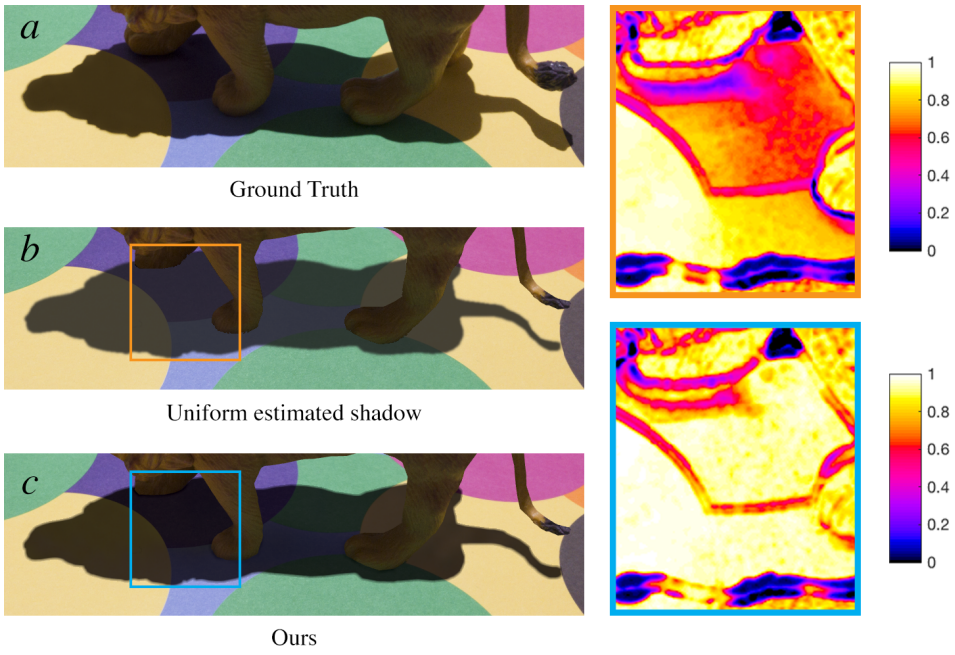


Fig. 3.11 (**a**) Ground-truth shadow generated by a physical occluder. (**b**) Uniform simple estimated shadow. (**c**) Virtual shadow rendered using our method. SSIM comparisons of ground truth with (b) above and (c) below showing improved accuracy through shadow retargeting.

show synthetic results generated using [Unity, 2019]. These show the behaviour of our technique under single and multiple points of light from point- to area-emitting regions. Our method is able to deliver accurate retargeting under a variety of lighting conditions. SSIM comparisons show high accuracy on umbra and penumbra regions between ground truth and our retargeting approach.

Figure 3.12 shows comparisons between ground-truth shadows and our retargeting algorithm in complex settings. These results were captured using real-time AR on an Apple iPhone X with [Vuforia, 2019] and [ARKit, 2019] frameworks. *Results (a–c)* demonstrate the algorithm's capability to handle an occluder placed distantly from the ground plane. *Results (d–f)* show the adeptness of the method to handle light sources casting over multiple-receiver surfaces. Our algorithm uses SLAM techniques to detect the ground and vertical planes using [ARKit, 2019]. SSIM results demonstrate good precision using our real shadow-retargeting

Task	Time	Percentage
Object Registration	1.83 ms	4.43 %
Auxiliary masks	2.59 ms	6.27 %
Deformable Object Retargeting	1.93 ms	4.67 %
Shadow Warping	3.13 ms	7.57 %
Discretized Search	15.78 ms	38.20 %
Uniform Shadow Blurring	7.36 ms	17.82 %
Percentage Closer Soft Shadows	5.27 ms	12.76 %
Background Inpainting	2.43 ms	5.88 %
Scene Rendering	0.98 ms	2.37 %
Total	**41.3 ms**	**100 %**

Table 3.1 Time breakdown of a typical frame processed using *Deformable Object Retargeting* and *Shadow Retargeting* techniques on an Apple iPhone X.

technique, whose accuracy may be further improved with 3D marker-less object registration and tracking.

Real-Time Performance

From figure 3.10, *results (a–d)* were generated with an Apple iPhone X with an output resolution of 2436 px by 1125 px. These achieved interactive frame rates, on average, 25 fps in a scenario with a fixed point of view, and 20 fps, in one with camera movement. *Results (e–j)* were generated in a 2.8 GHz Quad-core Intel Core i7 with 16 GB of RAM with an image size of 1280 by 720 px. These achieved a constant frame rate of 30 fps, which allowed us to reconstruct shadows in real time for modern video standards.

Table 3.1 breaks down the processing time of our technique under a fixed point of view setting. The primary bottleneck of our system is the discretized search for shadow reconstruction. On average, this takes one-third of the render time per frame. Nonetheless, the number of sample-search iterations is typically low, which makes the method suitable for interactive frame rates on low-powered mobile devices. Our approach proves to be consistent in rendering performance under different lighting scenarios.

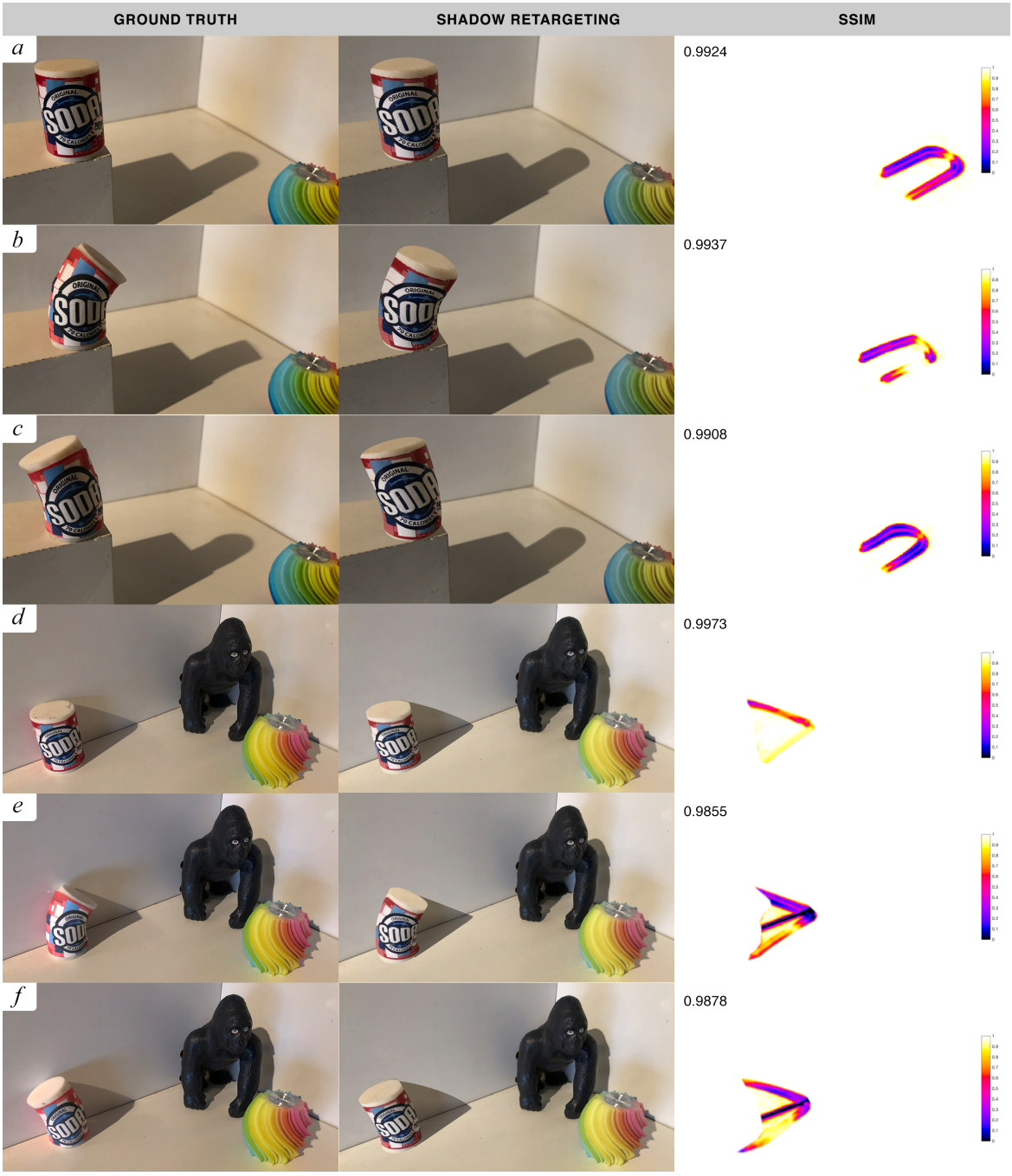


Fig. 3.12 Comparisons of ground truth with our shadow-retargeting scheme in complex settings. (**a**–**c**) Real 3D-printed nondeformed-object shadow retargeted with an occluder placed distantly from the ground plane. (**d**–**f**) Real 3D-printed non deformed-object shadow retargeted with a unique light source casting over multiple-receiver surfaces.

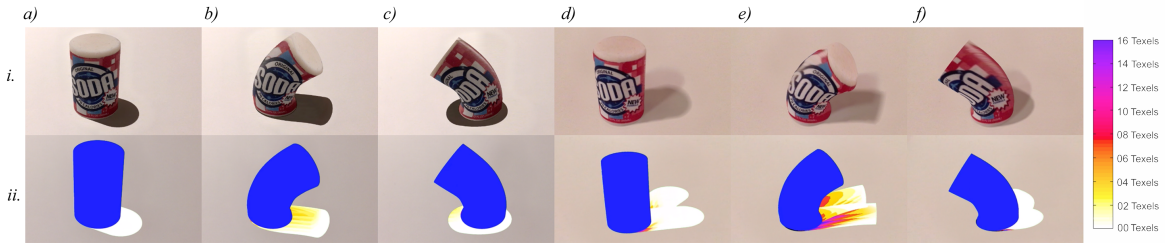


Fig. 3.13 **(Row i)** Reconstructed shadows at a retargeted pose. **(Row ii)** Color-encoded visualisations between distance from the occluded to the sampled texel, used as shadow probe. Results (**a**–**c**) were generated under a single light source casting hard shadows. Results (**d**–**f**) were generated under multiple light sources casting soft shadows.

Reconstruction Visualizations

As detailed in section 3.5.3, when an occlusion is present in a region of the warped shadow, we perform shadow inpainting using a discretized concentric ring search for appearance sampling. Our method selects the closest non-occluded texel. We used the binary mask from figure 3.7c for segmenting the area in which probe sampling is appropriate. When multiple shadows are present, this is done for each of them being classified by isolated and overlapping regions. Figure 3.13 displays the texel distance from sampling point for single and multiple light sources in hard- and soft-shadow scenarios. As can be observed in this figure, sampling distance is typically low, which leads to interactive frame rates, as presented in table 3.1.

3.6 Discussion

The proposed solution for light estimation covers a more flexible and adaptable method than previous approaches as very little configuration is required. It provides an unobtrusive and more user-friendly approach to estimate ambient lighting in an augmented-reality scenario. Nonetheless, the method requires flat Fresnel lens to be visible within the frustum of the camera. If these are not visible, the last known light direction is used from [ARKit, 2019] until lenses become visible by the camera again, which is often satisfactory for stable lighting

conditions. Further, our approach relies on light sources to be visible as reflected in the lens for each illumination environment sample. If these light sources would become occluded, our approach would accordingly update the estimation. As such, the method is most appropriate given a distant illumination environment assumption

Shadow Retargeting is the first approach to retarget already present shadows from static objects against their overlaid movement with plausible coherent results. The reference-point selection for the discretized concentric ring search algorithm could be further improved by sampling scene visibility. This would better estimate reconstructed shadow appearance. Additionally, this would allow for more advanced techniques to interpolate and even extrapolate the shadow to obtain more consistent results in even more complex scenarios. An approach based on bidirectional re-projection similar to [Yang et al., 2011] could resolve large deformations or topology changes in which occluder geometry is significantly altered from its physical position. The impact of such schemes on mobile real-time performance remains uncertain.

Our method holds plausibly for small-proximity displacements. Deformations made under the same visual axis are perceptually more precise. In the case of animated meshes that reveal a geometry not present in the reference model, visual artefacts appear due to dis-occlusions, addressed in part by inpainting. Temporary inconsistencies between frames of animation may appear due to the reconstruction search performed in real time for occluded areas. These limitations could be overcome by using a machine-learning model that would contain accurate predictions for retargeted poses under large translations or deformations. Finally, when performing reconstruction in a multiple-shadow scenario, we relied on sampling from a physical overlapping region of the shadow to maintain its appearance. Therefore, our method would fail to reconstruct if the physical shadow had no overlapping region that can be used to sample. We anticipate that this limitation could be surpassed by calculating the approximate overlapping appearance using the two projected individual shadows.

3.7 Summary

This chapter introduced *Appearance Retargeting* as an approach to bring inanimate bodies and shadows of static real-world objects to life. The following contributions have been presented:

- we have introduced the *Props Alive Software Framework*. This framework provides the structure required to bring toys to life in mobile AR (see section 3.2).
- we have presented a novel light-estimation approach employing a flat light probe fabricated using a flat reflective Fresnel lens. Unlike three-dimensional sampling probes, such as reflective spheres, Fresnel lenses can be incorporated into any type of object, such as a book or product packaging, providing a seamless light-estimation approach for AR experiences (see section 3.3).
- we have described *Object Retargeting*. This method achieves an illusion of movement from the real-world object through image retargeting techniques using Augmented Reality (see section 3.4).
- we have presented *Shadow Retargeting*. This method is an efficient and focused approach in which already present shadows from static real objects in the scene are retargeted according to virtual overlaid augmented reality movement (see section 3.5).

Chapter 4 will detail how *Appearance Retargeting* can be used in a CMR system that allow collaboration between multiple remote participants using objects brought to life in AR.

Chapter 4

Intermediated Reality

Intermediated Reality proposes a CMR framework to stimulate mediated collaboration among distributed peers using objects brought to life in AR. Section 4.1 introduces this novel concept of collaboration and entertainment. Section 4.2 defines the tele-puppetry model of interaction. Section 4.3 analyses the media richness of IR. Section 4.4 introduces *ToyMeet*, a CMR system that allows interactions between multiple remote participants. Section 4.5 measures the usability of the proposed CMR framework by undertaking a *System Usability Scale (SUS)* questionnaire with end-users. Section 4.6 provides a discussion of the work presented in the chapter. Section 4.7 concludes with a summary of the chapter.

4.1 Introduction

Intermediated Reality is a CMR framework that allows multiple users to access the same shared MR environment from remote locations using objects brought to life in interactive AR. This approach aims not only to allow users to collaborate remotely in a novel way, but also to enhance creativity, imagination and interaction with inanimate objects of our daily lives. In this sense, a CMR framework allows a more natural approach to mediated social interaction, in which multiple users can collaborate and interact with each other through

digital displays in a shared space. Our *ToyMeet* practical demonstration focuses on mixing real and virtual spaces seamlessly in a remote shared context. By augmenting the camera feed with our reconstructed appearance of the object in a deformed shape, we perform the illusion of movement for real-world static objects, remotely. We highlight the main contributions of this chapter as follows:

- we introduce, Intermediated Reality, a tele-present augmented reality framework that enables mediated communication and collaboration for multiple users through the remote possession of toys brought to life.
- inspired by the [Shannon and Weaver, 1949] model, we define the tele-puppetry model of interaction in IR (see section 4.2).
- we describe how voice and facial expressions data are transmitted to the system's database server and reproduced remotely in AR using the receiver's physical toy (see section 4.4).
- we measure the usability of the proposed CMR framework by undertaking a SUS questionnaire with end-users (see section 4.5.1).

4.2 Tele-Puppetry Model of Interaction

Interaction is the action of exchanging information between two or more participants in order to transmit or receive information through a shared system of signs and semantic rules. The [Shannon and Weaver, 1949] model (see figure 4.1a) is specially designed to develop an effective communication between the sender and the receiver. It contains context, sender, message, medium, receiver and feedback as the key components of the model.

- *Context* is the situation in which the communication is developed. It is the set of circumstances that affect both the sender and receiver, and also determine the interpretation of the message.
- *Sender* is the person who transmits a message. This is encoded using a combination of words understandable to the receiver.
- *Message* is the information that is exchanged between the sender and the receiver.
- *Medium* is the channel through which the encoder will communicate the message. This can be printed, electronic or audible and depends on the nature of the message and the contextual factors of the environment.
- *Receiver* is the person who interprets the message. This is influenced by the context to which it is exposed when decoding the message.
- *Feedback* is the response or reaction of the receiver to a message. The communication becomes effective when a response is emitted.

Drawing from the [Shannon and Weaver, 1949] model of communication, we define the tele-puppetry model of interaction in AR (see figure 4.1b). In this approach, we use toys brought to life as a channel to send and receive information. This means that additional components are now present on the loop of interaction. These new elements, which we define as the *sender's intermediary* and the *receiver's intermediary*, are the responsible for embodying and emitting information using a physical toy brought to life. These components are the real-world representation of the sender or the receiver in a remote venue.

In the tele-puppetry model of interaction, the *context* translates into our *Tele-present Mixed Reality (TMR)* context. Distributed participants share the same MR environment as a tele-present CMR system as defined by [Billinghurst and Kato, 1999]. Further, this context is not only linked with the *sender* and the *receiver*, but also with the *sender's mediator*, as the

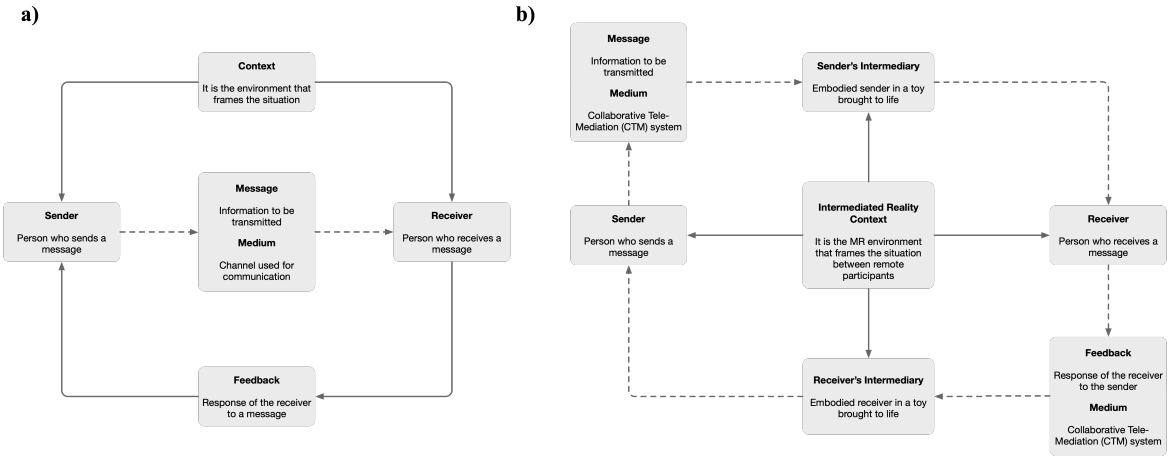


Fig. 4.1 **(a)** Shannon and Weaver model of communication. **(b)** Tele-Puppetry model of Interaction.

presenter of the emitted information from the sender in the receiver's location. Additionally, due to the distributed nature of a TMR system, the receiver's feedback to the sender is transmitted through the same system. In this case, the *receiver's mediator* physical toy presents the emitted information from the receiver in the sender's location operating the tele-puppetry model of interaction in the opposite direction. The previous receiver now becomes the sender, and the previous sender, now becomes the receiver.

The *mediator*, who emits the sender's information in a remote location, acts as a focus of the user-interaction for the receiver. Drawing from [Kennedy et al., 1996]'s framework for information visualization, which has low-latency as a key consideration being derived from a *Model–View–Controller (MVC)* architectural pattern, we present the tele-puppetry model of interaction as a *Model–View–Presenter (MVP)* architecture most similar to the definition of [Kennedy et al., 1996]. In this architectural pattern, the *model* is an interface defining the data to be displayed. The *view* is a passive interface that displays the data. The *presenter* acts upon the model and the view. It retrieves data from the model, and formats it for display in the view (see figure 4.2).

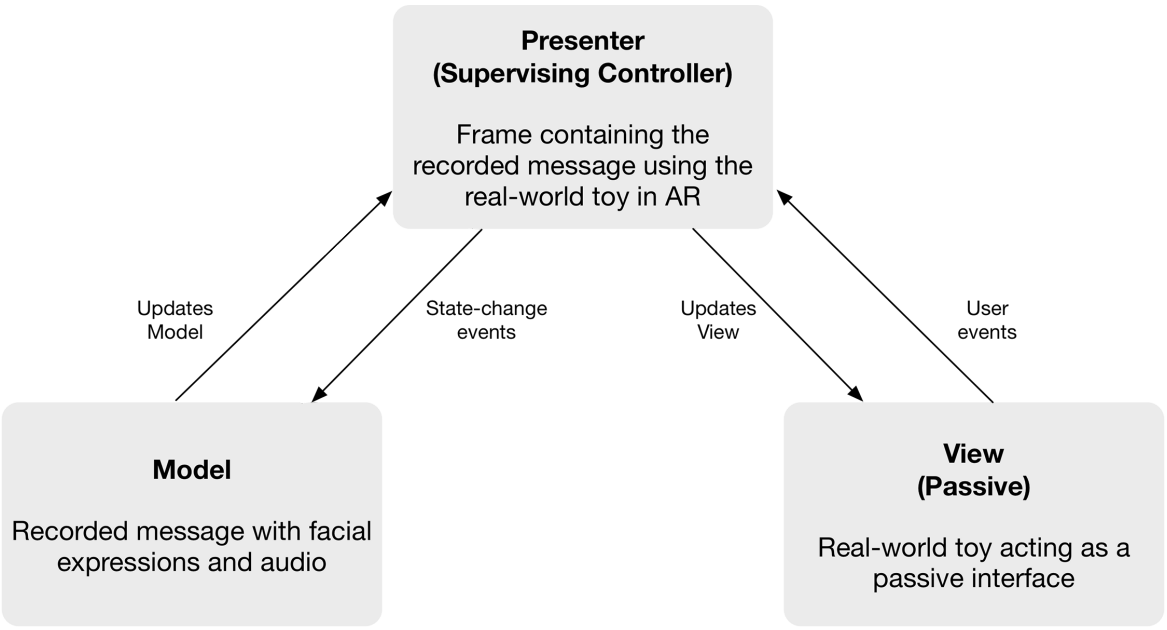


Fig. 4.2 The tele-puppetry model of interaction uses a *Model–View–Presenter (MVP)* architectural pattern. **Model** represents the recorded information with facial expressions and audio. **View** accounts for the registered real-world toy. **Presenter** is the rendered frame containing the recorded message using the real-world toy in AR.

4.3 Media Richness in Intermediated Reality

Fundamental in collaboration system design is *Media Richness Theory*, introduced by [Daft and Lengel, 1986], which we use to analyse the media richness of IR. Given the fact that audio, visual and facial cues can be reproduced on a remote intermediary, natural and body languages are seamlessly presented to the remote receiver.

With the aim of bringing IR experiences closer to a face-to-face experience, we use spatial audio from [Sodnik et al., 2006] to achieve high fidelity sound when transmitting audible signals. This approach, unlike traditional methods, attaches sound to a specific three-dimensional point in the space. As IR experiences rely on a physical puppet being part of the interaction, we embed the audio source in the upper mouth region of the puppet reproducing the message. As this audio is fully synchronised with the visual and natural cues reproduced on the real-world puppet, our low-latency framework is capable of achieving

Information Richness	Medium	Feedback	Channel	Source	Language
High	Face-To-Face	Immediate	Visual, Audio	Personal	Body, Natural
	Intermediated Reality	Fast	Visual, Audio	Personal	Body, Natural
	Videoconference	Fast	Visual, Audio	Personal	Natural
	Telephone	Fast	Audio	Personal	Natural
	Written, Personal	Slow	Limited Visual	Personal	Natural
	Written, Formal	Very Slow	Limited Visual	Impersonal	Natural
Low	Numeric, Formal	Very Slow	Limited Visual	Impersonal	Numeric

Fig. 4.3 Media richness theory that compares the level of media richness in IR with other types of media.

high medium richness and puts our system one step closer to the pursued instant telepresence (see figure 4.3).

4.4 ToyMeet

ToyMeet is a CMR system that allows interactions between multiple remote participants using toys brought to life in AR. In order to achieve this, we use *Object Deformable Retargeting* (see section 3.4) and *Shadow Retargeting* (see section 3.5) to bring to life the inanimate body and shadow of static real-world objects. The sender broadcasts voice and facial expression data to the system's database server (see section 4.4.1) and the receiver reproduces the AR content remotely using the receiver's physical toy (see section 4.4.2).

4.4.1 Capturing Sender's Information

In order to allow remote interactions between multiple participants, we first need to capture the information being emitted by the sender. To do so, this section describes the procedure employed to capture the user's voice and facial expressions.

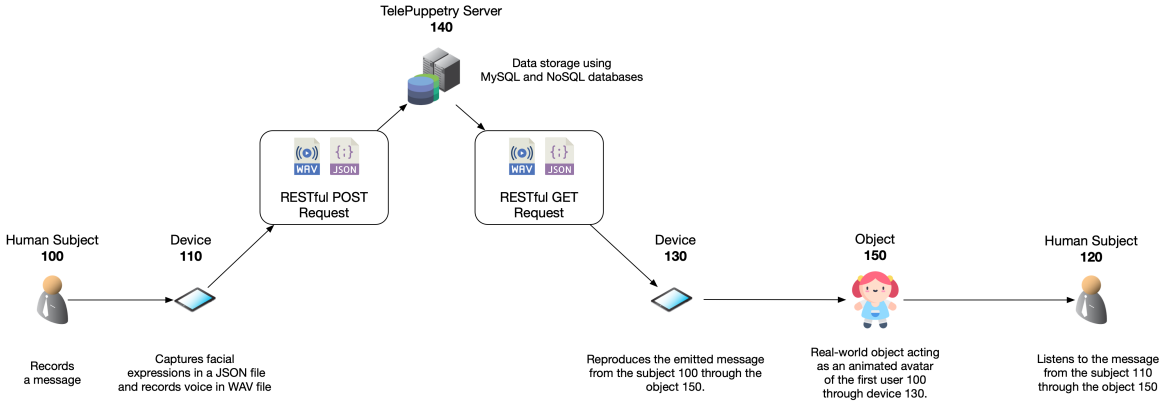


Fig. 4.4 *ToyMeet* system diagram. The sender broadcasts voice and facial expression data to the system's database server and the receiver reproduces the AR content remotely using the receiver's physical toy.

Recording Sender's Voice

The sender's voice is recorded with the microphone of their own mobile device. We initialize each sentence recording when the user taps the screen. Once this is tapped for a second time, we finalize the recording. The captured audio is buffered locally on the sender's mobile device and broadcasted to the server of the TMR system once the recording is finished (see section 4.4.3). Once the file has been broadcasted, the audio buffer is erased from the user's device. We encode the recorded voice using a stereo, 16-bit non compressed Waveform audio file format (*wav*) at 44.100Hz.

Acquiring Sender's Facial Expressions

To acquire the sender's facial expressions, we use a depth-enabled mobile phone that extracts facial features using [ARKit, 2019]. We initialize a recording session when the user taps the button displayed on the screen simultaneously with the voice. For every frame in which the recording session is active, we store the normalized weights of their voice phonemes and facial features for the complete list of attributes. This information is stored locally on the sender's mobile device and broadcast to the server of the IR system once the recording is

finished. Once the data has been broadcast, this animation data buffer is erased from the user's device.

In order to store the facial blend-shapes sequentially, we serialize their values using the *JavaScript Object Notation (JSON)* file format. When the session is initialized, we allocate a dynamic-sized array to memory. This array gets pushed with a new element on a per-frame basis. Each element of this array is a dictionary that contains the normalized values of each captured blend-shape.

4.4.2 Playing Messages on Objects Brought to Life

In order to use physical toys as a channel for tele-puppetry, we need to reproduce the information captured by the sender on them. This section describes how puppet facial expressions are created and correctly synchronized with the audio on playback time.

Puppet Facial Expressions

Our method for bringing puppets to life through AR requires a mesh for texture deformation. In order to recreate the facial expressions of the sender, we need to reproduce the captured blend-shapes into the puppet's mesh. These consist of 52 key voice phoneme and facial features. To do so, we use a standard 3D editor, such as Autodesk Maya, to create an animated mesh with key blend shapes. Each blend shape is normalized and weighted automatically accordingly with the registered data from the sender.

Adaptive Lip Syncing

Pulse Code Modulation (PCM) is a method used to digitally represent sampled analogue signals. In a PCM transmission, the amplitude of the analogue signal is sampled regularly at uniform intervals, and each sample is quantized to the nearest value within a range of digital steps. The levels of quantification vary according to the wave amplitude in PCM encodings.

To record the user's voice, we use the Waveform audio file format, more commonly known as *wav*, due to its uncompressed audio encoding. It uses *Linear Pulse Code Modulation (LPCM)* to linearly distribute the quantization levels across an audio transmission. This linearity allows us to adaptively synchronize the correct facial expression at a given time according to the current LPCM sample. This synchronization is possible because we know the total number of frames recorded for facial expressions and these coincide exactly with the duration of the audio. Such adaptive synchronization is of great importance when the frame rate of the reproduction device differs from the capturing hardware or when the rendering rate fluctuates and does not become constant. This approach does not produce any delay on audio and visual cues as these are always matched to the number of LPCM samples at a current given time.

To acquire the correct facial expression data (d) for any given modulation, we calculate equation 4.1 in real-time for each frame. ($s[t]$) is the number of LPCM samples at a time t of the audio clip. (s) is the total duration of the audio clip in LPCM samples. (n) is the total number of recorded frames that contain facial expressions.

$$d = \frac{s[t]}{\frac{s}{n}} \tag{4.1}$$

4.4.3 Performance Broadcast

To optimize for a low-latency communication, we broadcast serialized facial blend-shapes and recorded audio data in a single server request. Our framework streams the content using binary blobs of data in the form of a byte array. This data stream consists of concatenated bytes from the JSON and WAV file using an XML structure. The binary stream of data is transmitted to the server through a web socket that reads chunks of 8192 bytes at a time. Reading from the streaming continues until the file pointer has either reached the end of file

or read the entire byte length. The read data is written to the server's disk and labelled using the current time-stamp and user id.

With the aim to account for scenarios in which more than two participants interact or communicate with each other, our framework supports broadcasting of audio and visual cues to multiple participants. In this case, one participant of the session takes the host role, and all the other ones, subscribe to the session created by that participant. Interactions between peers are labelled, ordered and queued using timestamps, allowing a sequential and natural interaction between participants. This enables group chats and multiple collaborative interactions.

4.5 Experimental Assessment

Our CMR system targets mobile devices using interactive frame rates and low-latency interaction. Section 4.5.1 measures the usability of the proposed CMR framework by undertaking a SUS questionnaire with end-users. Section 4.5.2 analyses the rendering performance in an Apple iPhone X. Section 4.5.3 describes the system latency for miscellaneous mobile broadbands.

4.5.1 SUS Questionnaire

As Intermediated Reality proposes a new model of interaction not previously presented elsewhere, we understood the urge of analyzing the usability of our system from a Human-Computer Interaction point of view.

To evaluate our system, we made use of the SUS scale introduced by [Brooke, 1996]. The SUS scale provides a quick, reliable tool for measuring the usability of a wide variety of products and services. It consists of a 10 item questionnaire with five response options

for respondents; from *Strongly Agree* to *Strongly Disagree*. A SUS score above a 68 is considered to be above average.

The experiment took place in a fully empty room of around nine square meters. The room was well lit by two lights placed on the ceiling. The IR object, a plaster model pirate head, was placed on the wall at the height of one meter from the floor. Participants were initially briefed outside the room. They were told that they will be interacting with the plaster model pirate head through the mobile device and that once they enter the room, they will need to point towards the pirate head to start the experience.

Once participants entered the room and pointed towards the pirate head through their mobile device, they were prompted the following question: *"I have lost my treasure, could you help me out?"*. From here on, an open mediated conversation for the duration of two minutes took place for each participant. After two minutes, the experiment was ended and the participant was asked to fill out the SUS questionnaire. We received feedback from the ten participants, 8 male and 2 female, between the age of 22 and 39.

Figure 4.5 shows the results of the SUS questionnaire carried out by ten participants. From the results obtained, we can draw the conclusion that users would like to use an IR system frequently. Participants evaluated the framework as easy to use, without many things needed to be learned before being able to use it. However, due to the novel nature of the system, users did not feel totally confident with the system. Such feedback makes us understand that an in-app step-by-step tutorial is needed in order to make the user familiar with an IR system. Nonetheless, they affirmed that most people would be able to learn how to use the system once the IR interaction foundations are in place. Feedback provided by users state that the various functions of the system are well integrated and that IR is an engaging experience. Scores denote a positive outcome, with margin of improvement on the system usability to achieve an even more engaging experience with final end-users.

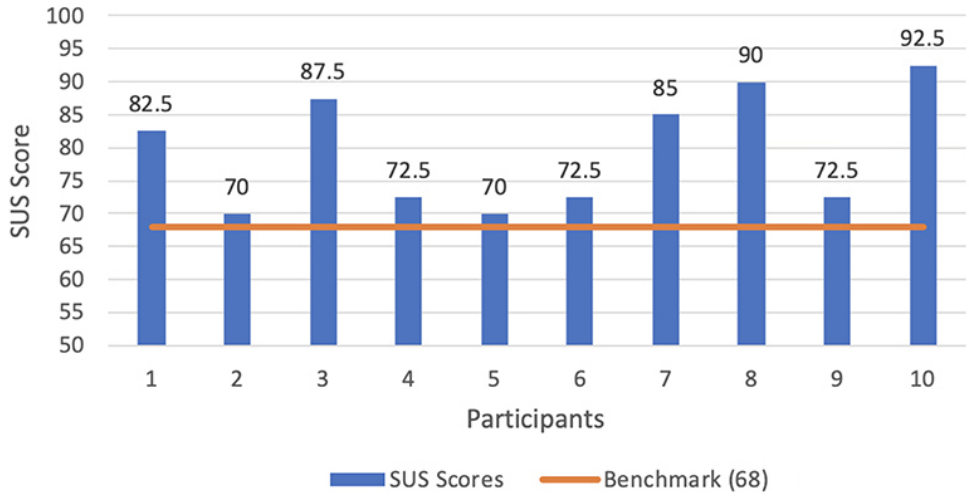


Fig. 4.5 Results of the SUS questionnaire on ten end-users using IR.

4.5.2 Rendering Performance

Our experimental assessment was performed using an Apple iPhone X with an output resolution of 2436px by 1125px. We achieved interactive frame rates on this mobile phone.

As seen in table 4.1, the primary bottleneck of our system is the *Shadow Retargeting* algorithm. As detailed in [Casas et al., 2018b], the sampling search for shadow reconstruction requires around *15ms* to achieve coherent results. In order to further optimize this sampling search, the reference-point selection for the discretized concentric ring search algorithm could be further improved by sampling scene visibility. This would better estimate reconstructed shadow appearance. Additionally, this would allow for more advanced techniques to interpolate and even extrapolate the shadow to obtain more consistent results in even more complex scenarios. An approach based on bidirectional re-projection similar to [Yang et al., 2011] could resolve large deformations or topology changes in which occluder geometry is significantly altered from its physical position. The impact of such schemes on mobile real-time performance remains uncertain. The rest of time invested in the *Shadow Retargeting* algorithm is for shadow warping, auxiliary masks and soft shadows. On average, retargeting the shape of the shadow takes approximately two thirds of the render time per frame.

Task	Time (ms)	Percentage (%)
AR Marker-less Tracking	1.85	3.91
Object Retargeting	4.87	10.31
Shadow Retargeting	34.13	72.26
Appearance Reconstruction	2.64	5.59
Background Inpainting	2.78	5.88
Scene Rendering	0.96	2.03
Total	**46.9**	**100**

Table 4.1 Time breakdown of a typical frame processed using *ToyMeet* in an Apple iPhone X.

The remainder third of the render time per frame is invested in miscellaneous duties. The second substantial task, which takes around 10% of the rendering time per frame, is object retargeting. This section encapsulates duties such as transferring texture data from the camera feed, rendering the overlaid mesh in a deformed position or assigning the weighted blend-shapes in real-time. Following this task, with an approximate 5% of the rendering time per duty, appearance reconstruction for occluded areas and background inpainting over uniform backdrops rank. Finally, marker-less AR tracking and scene rendering take the remainder 5% of the rendering time per frame.

4.5.3 System Latency

Our framework is optimized for low-latency communications among participants. As *ToyMeet* is designed to work efficiently in mobile devices, we analyse the broadcasting times in different broadbands. In table 4.2, we breakdown the data size per frame and second. Each serialized JSON blend-shape frame takes the size of 2,045 bytes. This includes the normalized weighted values of the key phonemes and facial features. In addition to the blend shapes, we record the syncronized audio using a stereo, 16-bit non-compressed WAV recording at 44,100 KHz. This has a bit-rate of 1,411.2 Kbps, which sizes at 5,880 bytes per frame. The combined captured data amount is 7,925 bytes per frame.

File Type	Bytes/Frame	KB/Second
JSON Data	2,045	61.35
WAV Audio	5,880	176.4
Total	**7,925**	**237.75**

Table 4.2 File size breakdown analysed per frame and second. Captured frame-rate is calculated at 30 fps. Recorded audio files are stereo, 16-bit at 44,100 KHz.

Broadband	Speed (Mbit/s)	Frame (s)
GPRS (2.5G)	0.115	0.55116
EDGE (2.75G)	0.237	0.26743
HSPA (3G)	5.8	0.01090
LTE (4G)	50	0.00126
WiFi	100	0.00063
eMBB (5G)	10,000	0.000063

Table 4.3 Time breakdown for broadcasting combined recorded audio and serialized blend-shapes in miscellaneous mobile broadbands. Calculated times do not take into account accidental lost packages caused by the user's environment, such as packet collision, radio interference or over-demanded service.

7,925 bytes per frame may seem like a small number, but when done at 30fps in a slow broadband, the transmission time can be a challenge. As it can be seen in table 4.3, this is specially the case for GPRS (2.5G) and EDGE (2.75G) connections, in which a sample message of 10 seconds could take almost 3 minutes to be broadcasted. This is not the case for faster connections, such as HSPA (3G) or LTE (4G). In this case, data transmissions are well optimized and broadcasting times are as little as 0.38 seconds for a sample message of 10 seconds in a LTE broadband. With the upcoming plans for 5G connections, with guaranteed minimum speeds of 10GBps, our framework will accomplish remote synchronous real-time capabilities. Hence, we understand that for a smooth and low-latency communication the user should have at least a HSPA (3G) broadband.

4.6 Discussion

Our system requires having a three-dimensional version of the real-world object with all blend-shapes modelled in order to create AR toy figures. This task is not simple and needs specific adjustments for each model, which requires an artist to manually modify each expression for each model. We anticipate that we will be able to reduce the time dedicated to this task by creating a set of predefined expressions that can be retargeted to new models by transferring the weights of the facial bones in the 3D model. This would set the foundations for a potential auto-rigging and skinning system that would speed up the animation process of organic and non-organic objects. Such approach would reduce the time needed to create blend-shapes, since the only adjustment required would be to fine-tune parameters according to each 3D model.

In applying our concept more broadly, we foresee applications of our framework across the entire RVC. Using full immersive VR, in which the physical intermediary is not seen, we anticipate the use of haptics systems and virtual characters for getting in touch with the receiver's physical space similar to [Bierz et al., 2005]. In a robotic telepresence scenario, in which the intermediary is a robot with spatial audio systems and physical animated facial expressions, we envision the robot to be driven by our system using synchronized audio and captured facial data following [Danev et al., 2017]. Hence, we understand that IR has a broad scope of applications in miscellaneous industrial sectors.

4.7 Summary

In this chapter, we have described Intermediated Reality, a CMR framework that allows multiple users to access the same shared MR environment from remote locations using objects brought to life in interactive AR. This approach aims not only to allow users to collaborate remotely in a novel way, but also to enhance creativity, imagination and interaction with

inanimate objects of our daily lives. In this sense, a CMR framework allows a more natural approach to mediated social interaction, in which multiple users can collaborate and interact with each other through digital displays in a shared space. The following contributions have been presented in this chapter:

- we have introduced, Intermediated Reality, a tele-present augmented reality framework that enables mediated communication and collaboration for multiple users through the remote possession of toys brought to life.
- inspired by the [Shannon and Weaver, 1949] model, we defined the tele-puppetry model of interaction in IR (see section 4.2).
- we described how voice and facial expressions data are transmitted to the system's database server and reproduced remotely in AR using the receiver's physical toy (see section 4.4).
- we measured the usability of the proposed CMR framework by undertaking a SUS questionnaire with end-users (see section 4.5.1).

Chapter 5 will detail how IR can be used in a broad range of applications across the RVC.

Chapter 5

Reality Mixer

Intermediated Reality is designed to account for a diversity of MR applications in which a collaborative environment can be beneficial for distributed participants. Section 5.1 proposes IR as a method of communication and entertainment through toys brought to life in mobile AR. Section 5.2 presents an innovative form of gaming that encompasses interactions within the RVC continuum. Section 5.3 presents the benefits of applying IR to Stop Motion animation. Section 5.4 provides a discussion of the work presented in the chapter. Section 5.5 concludes with a summary of the chapter.

5.1 Distributed Communication

Our CMR system focuses on mixing real and virtual spaces seamlessly in a remote shared context. By augmenting the camera feed with the reconstructed appearance of the object in a deformed shape, we perform the illusion of movement for real-world static objects, remotely. As part of a two-way conversation, each person communicates through a toy figurine that is remotely located in front of the other participant. Each person's face is tracked through the front camera of their mobile devices and the tracking pose information is transmitted to the remote participant's device along with the synchronized voice audio, allowing an interactive

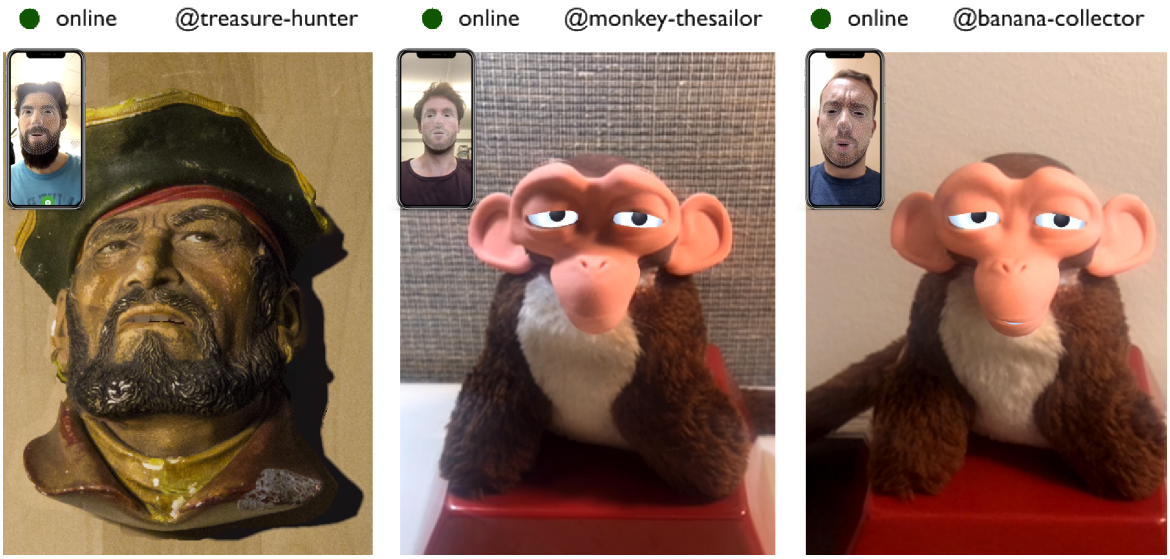


Fig. 5.1 Multiple participants using IR as a method of collaboration and entertainment through the remote possession of toys that come to life in AR.

avatar chat (see figure 5.1). Up to this present time, no previous system has been documented to provide telecommunication via object animation in AR. This enables tele-puppetry to a broad range of applications, including for instance, group chats and collaborative interactions.

Besides telepresence, we foresee applications of our framework for remote tele-parenting. By being in the moment, spending quality time and showing warmth, care and respect, the relationship with the child can be strengthening. However, due to the commitments of adults, sometimes parents must be absent for a certain period of time. When this happens, telepresence aims to reduce physical distance by giving the feeling of being in that other location through remote movements, actions or voice. Our technique can make use of augmented reality through traditional toys to reproduce recorded messages of close relatives. We speculate that this would make parents and children virtually closer, when in reality, both are far away from each other in the real world. Each of them interacts with other participants through animated physical puppets, helping to awaken the imagination of the child and improving the ability to socially interact with others. Further, we foresee our IR framework

as a tool for compelling storytelling using toys brought to life in AR. We propose to embody the participant into a real-world toy as the narrator of a compelling story.

5.2 Multi-Reality Games

Interactive play can take very different forms, from playing with physical board games to fully digital video games. In recent years, new video game paradigms were introduced to connect real-world objects to virtual game characters. However, even these applications focus on a specific section of the RVC, where the visual embodiment of characters is either largely static toys in the real world or pre-animated within the virtual world according to a determined set of motions.

We introduce a novel concept, called Multi Reality Games, that encompasses interactions with real and virtual objects to span the entire spectrum of the RVC, from the real world to digital and/or back. Our application on real-virtual game interaction makes an evolutionary step toward the convergence of real and virtual game characters. Rather than static toys or pre-built and unconfigurable virtual counterparts, we bring together technologies from the entire RVC to target new game experiences.

We showcase our framework by proposing a game application on a mobile device. Without the need to change the location or set, we enable intuitive and seamless interactions between physical, augmented and virtual elements. The experience brings both worlds closer, and enables the user to customize the virtual scenario according to physical references. Section 5.2.1 presents the background knowledge of this novel form of play. Section 5.2.2 describes our approach to Multi Reality Games. Section 5.2.3 demonstrates our gaming experience using a mobile game implementation.

Fig. 5.2 **(a)** Static and inanimate 3D printed narrator. **(b-c)** Narrator speaking to the user through photo-realistic AR using a mobile phone.

5.2.1 Background

With the ever growing trend towards virtualization of everyday life experiences, MR systems have become a main area of interest, both in research and in the industry. Such systems are applicable to a wide range of areas, including the automotive sector, surgery, office environments and entertainment. To enable convincing immersive experiences, a seamless inclusion of the virtual content is required. Moreover, many applications require real-time interactions on low-powered devices, such as smart-phones or tablets, which due to their limited computational resources demand refined implementations of efficient algorithms.

Recently, significant advances have been made in the development of MR applications, gradually enhancing interactions between virtual and real content. However, if many of these technologies find a relevant place in the RVC, seamlessly transitioning between them — and therefore spanning the entire continuum within a single application — remains an open challenge.

Our research introduces a concept, that we call Multi Reality Games, that encompasses interactions with real and virtual objects throughout the spectrum of the RVC. We bring together some of the latest technologies in 3D scanning, object augmentation and character control to build a diversified and engaging application. The user is seamlessly driven through the entire RVC, which enables a progressive immersion into the virtual world, as well as a variety of interactions and customizations.

5.2.2 Approach

We propose Multi-Reality experiences as a natural, fluid and seamless approach to travel throughout the RVC. In this section, we detail the interest of going from reality to virtuality. Nonetheless, this could also be experienced in the opposite direction, from virtuality to reality.

Real-world games, such as puzzles or quizzes, help children learn how to interact with others and develop skills that are essential for life. Active play helps them with cognitive, creative and communicative skills. Our approach proposes combining those benefits with the advantages of virtual gaming through mobile devices. We enable the user to interact with physical objects anywhere in the real world. These can be assembled manually from pieces, built from scratch, be fully customized or even printed in 3D. This flexibility allows the user to improve their creativity while being able to better understand the foundations of game design.

Once the user has chosen the physical components necessary for the game and their position in the real world, we propose to enhance the game-play through a mobile device. Following the RVC, we present a progressive immersion from reality to virtuality. From our initial real-world, we transition to Augmented Reality, where virtual objects are now coexisting with the physical ones. The user can now interact with both real and virtual objects at the same time. Hence, physical items can be re-arranged at best convenience according to the recently overlaid virtual ones. As we propose a seamless blending within the entire RVC, we contemplate both worlds to interact with each other. For predefined physical objects, we even encompass enlivened real-world game assets. These are pre-registered targets that through the use of an animated mesh that has exactly the same shape as the physical object and the use of few image processing techniques, make the illusion of coming alive through the screen of the mobile device.

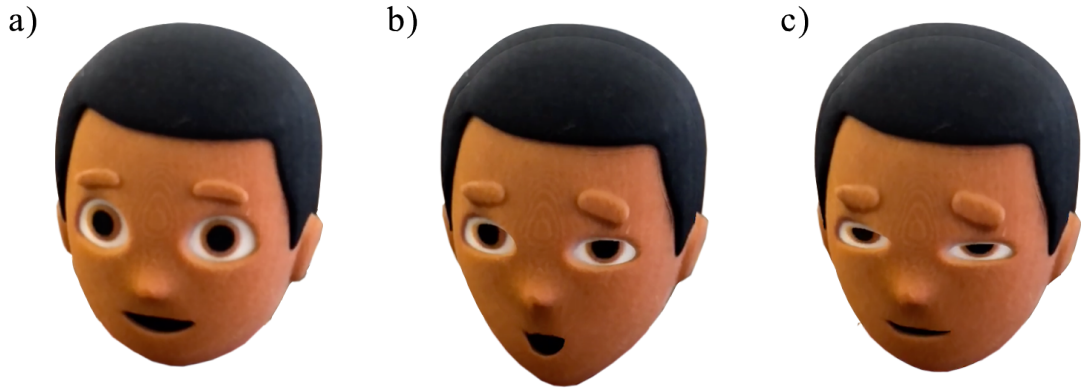


Fig. 5.3 **(a)** Static and inanimate 3D printed narrator. **(b-c)** Narrator speaking to the user through interactive AR using a mobile phone.

After some interaction with the MR environment, we propose a smooth transition to the completely virtual world. In this alteration of reality, the real world scenario vanishes and the environment becomes completely virtual. The user can now experience the game in a parallel reality that allows a Multi-Reality experience throughout the RVC.

5.2.3 Game experience

In our specific game experience, we propose an adventure game that travels throughout the RVC. The goal is to solve a quest to restore freedom in the hidden world of *Tasbada*. Please refer to the accompanying video for visualizing the complete game-play. Following our approach described in section 5.2.2, we enable the user to start the game in the physical world. We let the user decide where the Multi-Reality experience will take place and select the physical objects that will compose the real-world scene.

As soon as the user has defined a location and the real-world assets for the gaming experience, we start travelling across the RVC. We use a predefined 3D printed object to incorporate a narrator into the game (see figure 5.3). We use the *Appearance Retargeting* approach introduced in chapter 3 to create the illusion of movement of static objects from the real world with photo-realistic renderings.

Once the narrator has explained the mission to the player, we ask him to take a photo of himself (or a friend) to be transported into the game. Extending the work from [Nijholt, 2005] on enabling meetings in the virtuality continuum, we use virtual avatars for a gaming experience across the RVC. This brings closer the physical and augmented worlds as the user can feel connected to the virtual character. We use *AR Poser* from [Cimen et al., 2018] to capture the user's initial pose, and we additionally extract the representative color of the clothes to apply on the avatar.

Now that the player is embodied into the virtual character, he is asked to solve a quest by unlocking several milestones in the augmented world. We use *PuppetPhone* by [Anderagg et al., 2018] to let the user control his avatar with the movement of the phone. Once the milestones are accomplished and the final goal of the game is reached, we embark the user in a fully virtual world, as a reward for their achievement. Virtual objects present in the AR scene remain, but the real-world scenario becomes overlaid with a fully virtual setting. This last transition allowed the user to smoothly progress on step further in the RVC, from AR to VR. The user started the game at the real end of the continuum and ended it on the totally virtual side of it.

5.3 Stop Motion Animation

Stop motion animation techniques have an extensive history in the film industry. The first representative sample is, *"The Humpty Dumpty Circus"* by [Blackton and Smith, 1898]. This short animated film featured a circus of steady acrobats and animals in motion. The first Academy Award nomination for this genre came when [Noyes, 1964] redefined the technique of free-form clay animation and settled the foundations for modern stop motion pictures with the animated film *"Clay or the Origin of the Species"*.

Undoubtedly, one of the major exponents of this genre is *"The Nightmare Before Christmas"* produced by [Burton, 1993]. In this 76 minute long movie, 227 puppets were con-

structed, using more than 400 different expression heads for the main character. This movie was recorded at 24 frames per second, capturing a remarkable amount of 109,400 frames for its entire length, taking 3 years to produce.

Recently, *"Kubo and The Two Strings"* has introduced 3D printing advancements to mass produce puppets and faces, and consequently, accelerate the film production. However, over 66,000 faces were 3D printed, hand-detailed, post-produced, set and captured in order to be ready for capture. Therefore, regardless the advantages that 3D printing brings to stop motion, a tremendous amount of manual post-processing time remains.

Intermediated Reality introduces an approach that reduces the number of 3D printed puppet components required in a traditional stop motion animation. To do so, the in-between frames of the key poses are computer-generated, having the only requirement to have the key poses physically present. Section 5.3.1 introduces augmented stop motion animation in-betweening. Section 5.3.2 describes an approach for fast facial posing of physical puppets in augmented stop motion animation.

5.3.1 Augmented Stop Motion Animation In-betweening

Stop motion animation evolved in the early days of cinema with the aim to create an illusion of movement with static puppets posed manually on each frame. Current state-of-the-art stop motion movies are introducing 3D printing techniques to produce animations more rapidly and with less production cost. We extend this by using IR as a real-time interactive system capable of generating virtual in-between poses according to a reduced number of key frame physical props (see appendix A). We perform deformation of the surface camera samples to accomplish smooth animations with retained visual appearance and incorporate a diminished reality method to allow virtual deformations that would, otherwise, reveal undesired background behind the animated mesh (see figure 5.4).

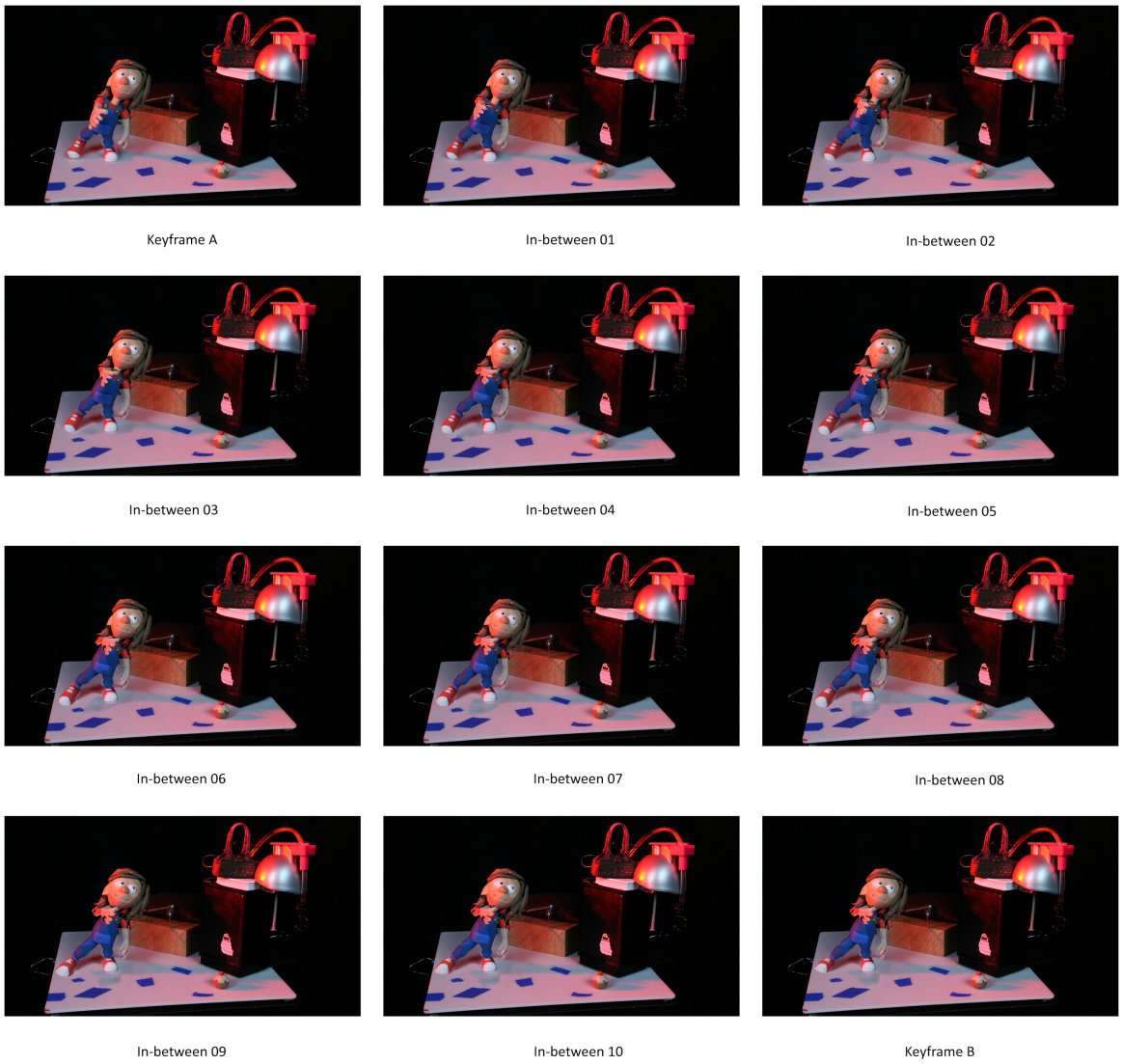


Fig. 5.4 Transition between keyframes A and B using *Augmented Stop Motion Animation In-betweening* to generate laborious in-between frames.

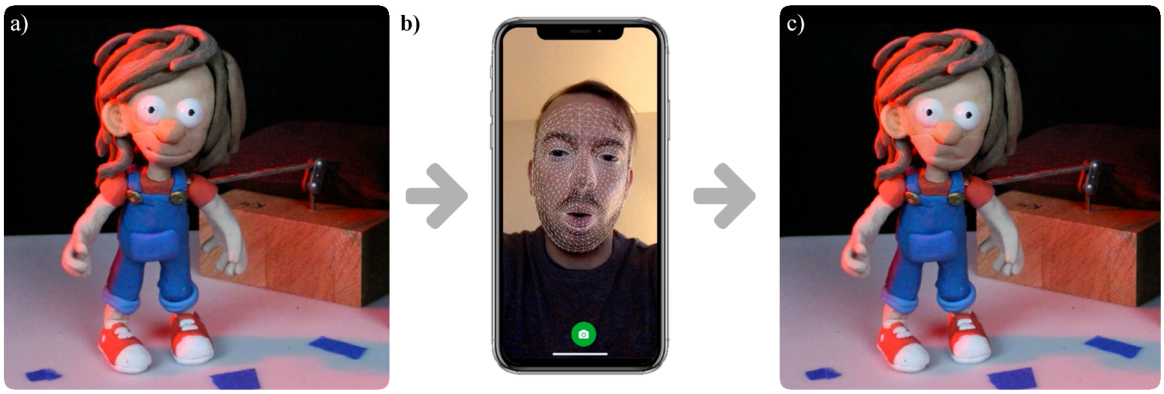


Fig. 5.5 **a)** Real-world stop motion puppet in an idle pose. **b)** User posing the facial expressions of a physical puppet using his own mobile phone. **c)** Directed open mouth in a real-world puppet using photo-realistic AR.

5.3.2 Fast Facial Posing of Physical Puppets in Stop Motion Animation

Accurately posing stop motion puppets for long-takes movies, frame by frame, is an analogue job that requires a high cost in resources and time. Recent approaches from [Casas et al., 2018b] or [Abdrashitov et al., 2018] aim to reduce these by generating and optimizing in-between frames digitally.

In this section, we introduce a technique in which we can direct a character's facial expressions directly from a mobile phone using Augmented Reality. We propose a method in which a 3D-printed puppet can be directed with the acquired blend-shapes from a user. We use [ARKit, 2019] to acquire the weighted values from the user and apply them to a rigged and skinned mesh that has the same exact shape as the printed one. This requires to create the 52 key voice phoneme and facial features for each character of a movie. This technique allows high fidelity with the desired result and an accurate synchrony of lips with the recorded audio (see figure 5.5).

5.4 Discussion

Participants who took part of our pilot study found IR an engaging experience overall. The novelty of seeing a tangible object come to life through the screen of their mobile device produced them a combined reaction of surprise and excitement. Few of these participants needed assistance to understand how IR worked, and when providing feedback, they advised to make use of a better user interface in the app to guide them in the process of locating the IR object.

The need for a precise calibration of the virtual object with the physical one to obtain a seamless IR experience was extensively discussed among collaborators. To solve such time-consuming calibration issues, we developed a calibration app with UI buttons to calibrate on-site the virtual mesh overlaying the physical object for a seamless IR experience.

In the case of stop motion animation applications, we got to learn that the technology produces high quality in-betweens, as praised by our art student collaborator, Kieran McLister. Nonetheless, he also remarked that it is not certain if our system would enable enough artistic freedom for animators to work unconstrained.

5.5 Summary

In this chapter we have described a diversity of MR applications in which IR can be beneficial for distributed participants. The following contributions have been presented in this chapter:

- Intermediated Reality has been leveraged to enable distributed remote communication among peers (see section 5.1).
- Multi-Reality Games have been presented. This approach is an innovative form of gaming that encompasses interactions with real and virtual objects throughout the spectrum of the RVC using IR (see section 5.2).

- an industrial application of IR has been demonstrated for Stop Motion Animation production work-flows. The approach is capable of creating realistic in-betweening for reduced production costs and can direct facial expressions of a character remotely from a mobile phone (see section 5.3).

Chapter 6

Conclusion

In this thesis we have introduced Intermediated Reality. Section 6.1 gives a summary of the contributions introduced in chapters 3, 4 and 5. Section 6.2 answers the research questions stated in chapter 1. Section 6.3 provides a discussion of our work and details future work in IR.

6.1 Summary

This thesis explored technical solutions to address the gap between virtual and physical worlds towards photo-realistic interactive AR. To convincingly re-animate physical objects through digital displays, a method of texture deformation with object inpainting was introduced. This technique, in combination with a method to retarget deformed virtual shadows and a solution to perform environment illumination estimation using inconspicuous flat Fresnel lenses, brought real-world props to life in a compelling and practical way.

Each method was integrated together to form Intermediated Reality, a tele-present AR framework that enables mediated communication and collaboration for multiple users through the remote possession of toys brought to life. Our framework presented applications for avatar chat communication, stop-motion animation movie industry and computer gaming

sectors. Concretely, an approach to reduce the number of physical configurations needed for a stop-motion animation movie by generating the in-between frames digitally in AR was demonstrated. AR-generated frames preserved its natural appearance and achieve smooth transitions between real-world key-frames and digitally generated in-betweens. Further, IR techniques extended across the entire RVC to target new game experiences called Multi-Reality games. This gaming experience made an evolutionary step toward the convergence of real and virtual game characters for visceral digital experiences.

6.2 Answers to Research Questions

- **RQ1**. *How are physical objects convincingly re-animated through digital displays using mobile AR?*

  Producing visually realistic compositions at interactive rates in AR is becoming of increasing importance in both research and commercial applications. With the aim to create convincing illusions of movement for otherwise inanimate objects, this thesis presented Object Deformable Retargeting. This method achieves an illusion of movement from real-world objects in AR (see section 3.4, [Casas et al., 2017]).

  - **RQ1.1**. *How are areas not present in the physical object reconstructed when revealed in AR?*

    For areas not present in the physical reference model, we are in need to reconstruct the revealed appearance in a plausible way. As our method uses a three-dimensional mesh that matches the physical object shape, we can accurately establish beforehand regions that will be revealed and are not visually seen in the reference puppet. This thesis has introduced appearance reconstruction, a method that is capable of reconstructing areas not present in the physical

reference model when revealed in AR (see section 3.4.2, [Casas and Mitchell, 2019]).

- **RQ2**. *How are physical shadows plausibly retargeted according to the animated movement of the real-world object in AR?*

  In order to retarget already present shadows from static real objects according to a virtual overlaid AR movement, this thesis has presented Shadow Retargeting. Our approach synthesizes virtual shadows directly by sampling the image that contains the shadow of the object presented in the real world. As the virtual mesh is animated, we warp the real shadow to retarget its appearance to the virtual one. Where the shadow warp would sample an invalid area, we reconstruct the appearance of the shadow by performing a discretized concentric ring search to find the closest valid sampling point (see section 3.5, [Casas et al., 2018b]).

  - **RQ2.1**. *How is the environmental illumination estimated employing a portable approach that seamlessly blends with everyday objects?*

    To accurately account for pose deformations in the scene, we are in need to estimate light source directions that cast shadows to the object that we intend to retarget. In this thesis, we have introduced a novel light-estimation approach employing a flat light probe fabricated using a flat reflective Fresnel lens. Unlike three-dimensional sampling probes, such as reflective spheres, Fresnel lens can be incorporated into any type of object, such as a book or product packaging, providing a seamless light-estimation approach for AR experiences (see section 3.3, [Casas et al., 2019]).

- **RQ3**. *How are objects brought to life integrated into a collaborative distributed environment?*

Objects brought to life in AR can be used to stimulate mediated collaboration among distributed peers by the use of Intermediated Reality, a CMR framework that allows multiple users to access the same shared MR environment from remote locations. This approach aims not only to allow users to collaborate remotely in a novel way, but also to enhance creativity, imagination and interaction with inanimate objects of our daily lives. In this sense, a CMR framework allows a more natural approach to mediated social interaction, in which multiple users can collaborate and interact with each other through digital displays in a shared space. Our practical demonstration focuses on mixing real and virtual spaces seamlessly in a remote shared context. By augmenting the camera feed with our reconstructed appearance of the object in a deformed shape, we perform the illusion of movement for real-world static objects, remotely (see chapter 4, [Casas and Mitchell, 2019]).

6.3 Future Work

Intermediated Reality presents an exciting venue for novel applications and methods of interaction in AR. Our vision is that this thesis is only the beginning of a line of research that can be extended and worked upon to solve several current open problems in AR.

We understand that the progress to be made in the field of marker-less tracking and SLAM, will further reduce calibration issues and enable even more seamless IR experiences across any physical object size or type. In this regard, objects which are currently hard to track, such as those with very few features, could potentially become trackable and the range of potential interactive objects much more extensive. We foresee that tracking might be reinforced by the use of AI, with special attention to specular and glossy materials, which are currently very hard to track. The use of machine learning could also solve the detection and understanding of deformable models. This would enable IR to work seamlessly in deformable plush toys, for example.

Our object and shadow retargeting solution holds plausibly for small-proximity displacements. Deformations made under the same visual axis are perceptually more precise. In the case of animated meshes that reveal a geometry not present in the reference model, visual artefacts appear due to dis-occlusions, addressed in part by inpainting. Large translations in the virtual object from the physical reference position cause lighting inconsistencies with spatial variations in the environment. Temporary inconsistencies between frames of animation may appear due to the reconstruction search performed in real time for occluded areas. We foresee that future wok using a machine learning model could solve for large translations or deformations.

In our current framework, having a mesh of the IR object is a requirement. However, we foresee that future advancements in MR technologies will contribute for even more seamless experiences of the concept we have developed. Concretely, if our system would be able to model and animate any object that is seen through the camera, this would enable more diverse experiences that could be adapted to any physical environment.

Further, we foresee that advancements in light-weight AR glasses or holographic technologies would make the distinction between physical and virtual elements even less discernible. In this case, the blending between real and virtual worlds would become imperceptible, which would enhance the user experience of our technology. In this regard, we foresee the use of IR on everyday objects using automatic object detection and animation.

References

[Abdollahi et al., 2017] Abdollahi, H., Mollahosseini, A., Lane, J. T., and Mahoor, M. H. (2017). A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 541–546.

[Abdrashitov et al., 2018] Abdrashitov, R., Jacobson, A., and Singh, K. (2018). f-stop: a system for 3d printed stop-motion facial animation. *Graphics Interface Posters, 2018*.

[Alamri et al., 2010] Alamri, A., Cha, J., and El Saddik, A. (2010). AR-REHAB: An augmented reality framework for poststroke-patient rehabilitation. *IEEE Transactions on Instrumentation and Measurement*, 59(10):2554–2563.

[Anderagg et al., 2018] Anderagg, R., Ciccone, L., and Sumner, R. W. (2018). Puppetphone: Puppeteering virtual characters using a smartphone. In *Proceedings of ACM SIGGRAPH Conference on Motion, Interaction and Games*.

[Andre, 2006] Andre, E. (2006). Engaging in a conversation with synthetic agents along the virtuality continuum. In *2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 19–20.

[ARKit, 2019] ARKit (2019). Apple ARKit. https://developer.apple.com/arkit/.

[Baba and Asada, 2003] Baba, M. and Asada, N. (2003). Shadow removal from a real picture. In *ACM SIGGRAPH 2003 Sketches &Amp; Applications*, SIGGRAPH '03, pages 1–1, New York, NY, USA. ACM.

[Basri and Jacobs, 2001] Basri, R. and Jacobs, D. (2001). Photometric stereo with general, unknown lighting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II.

[Bauer et al., 2001] Bauer, M., Bruegge, B., Klinker, G., MacWilliams, A., Reicher, T., Riz, S., Sandor, C., and Wagner, M. (2001). Design of a component-based augmented reality framework. *Proceedings - IEEE and ACM International Symposium on Augmented Reality, ISAR 2001*, pages 45–54.

[Beardsley et al., 2005] Beardsley, P., van Baar, J., Raskar, R., and Forlines, C. (2005). Interaction using a handheld projector. *IEEE Computer Graphics and Applications*, 25(1):39–43.

[Bertalmio et al., 2003] Bertalmio, M., Vese, L., Sapiro, G., and Osher, S. (2003). Simultaneous Structure and Texture Image Inpainting. *IEEE Transactions on Image Processing*.

[Besbes et al., 2012] Besbes, B., Collette, S. N., Tamaazousti, M., Bourgeois, S., and Gay-Bellile, V. (2012). An interactive augmented reality system: A prototype for industrial maintenance training applications. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 269–270.

[Bierz et al., 2005] Bierz, T., Dannenmann, P., Hergenrother, K., Bertram, M., Barthel, H., Scheuermann, G., and Hagen, H. (2005). Getting in touch with a cognitive character. In *First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics Conference*.

[Billinghurst and Kato, 1999] Billinghurst, M. and Kato, H. (1999). Collaborative Mixed Reality. In *Mixed Reality*, pages 261–284.

[Billinghurst et al., 1998] Billinghurst, M., Weghorst, S., and Furness, T. (1998). Shared space: An augmented reality approach for computer supported collaborative work. *Virtual Reality*, 3(1):25–36.

[Blackton and Smith, 1898] Blackton, J. S. and Smith, A. E. (1898). The humpty dumpty circus. *Kalem Company*.

[Boom et al., 2017] Boom, B. J., Orts-Escolano, S., Ning, X. X., McDonagh, S., Sandilands, P., and Fisher, R. B. (2017). Interactive light source position estimation for augmented reality with an rgb-d camera. *Computer Animation and Virtual Worlds*, 28(1):e1686. e1686 cav.1686.

[Brooke, 1996] Brooke, J. (1996). Sus: A quick and dirty usability scale.

[Burton, 1993] Burton, T. (1993). The nightmare before christmas. *Walt Disney Pictures*.

[Calian et al., 2013] Calian, D., Mitchell, K., Nowrouzezahrai, D., and Kautz, J. (2013). The Shading Probe: Fast Appearance Acquisition for Mobile AR. *SIGGRAPH Asia 2013 Technical Briefs on - SA '13*, pages 1–4.

[Calian et al., 2018] Calian, D. A., Lalonde, J.-F., Gotardo, P., Simon, T., Matthews, I., and Mitchell, K. (2018). From faces to outdoor light probes. *Computer Graphics Forum*, 37(2):51–61.

[Casas, 2019] Casas, L. (2019). Intermediated reality. In *SIGGRAPH Asia 2019 Doctoral Consortium*, SA '19, New York, NY, USA. Association for Computing Machinery.

[Casas et al., 2018a] Casas, L., Ciccone, L., Çimen, G., Wiedemann, P., Fauconneau, M., Sumner, R. W., and Mitchell, K. (2018a). Multi-reality games: An experience across the entire reality-virtuality continuum. In *Proceedings of the 16th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '18, pages 18:1–18:4, New York, NY, USA. ACM.

[Casas et al., 2018b] Casas, L., Fauconneau, M., Kosek, M., , Mclister, K., and Mitchell, K. (2018b). Image based proximate shadow retargeting. In *Proceedings of the Computer Graphics and Visual Computing (CGVC) Conference 2018*, Swansea, Wales, United Kingdom.

[Casas et al., 2019] Casas, L., Fauconneau, M., Kosek, M., Mclister, K., and Mitchell, K. (2019). Enhanced shadow retargeting with light-source estimation using flat fresnel lenses. *Computers*, 8(2).

[Casas et al., 2017] Casas, L., Kosek, M., and Mitchell, K. (2017). Props Alive : A Framework for Augmented Reality Stop Motion Animation. In *2017 IEEE 10th Workshop on Software Engineering and Architectures for Realtime Interactive Systems (SEARIS)*, Los Angeles, California, USA.

[Casas and Mitchell, 2019] Casas, L. and Mitchell, K. (2019). Intermediated reality: A framework for communication through tele-puppetry. *Frontiers in Robotics and AI*, 6:60.

[Castro et al., 2012] Castro, T. K. D., Figueiredo, L. H. D., and Velho, L. (2012). Realistic shadows for mobile augmented reality. *Proceedings - 2012 14th Symposium on Virtual and Augmented Reality, SVR 2012*, pages 36–45.

[Chen et al., 2007] Chen, J., Paris, S., and Durand, F. (2007). Real-time edge-aware image processing with the bilateral grid. *ACM Trans. Graph.*, 26(3).

[Cimen et al., 2018] Cimen, G., Maurhofer, C., Sumner, B., and Guay, M. (2018). AR Poser: Automatically Augmenting Mobile Pictures with Digital Avatars Imitating Poses. In *12th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing 2018*.

[Comport et al., 2006] Comport, A. I., Marchand, E., Pressigout, M., and Chaumette, F. (2006). Real-time markerless tracking for augmented reality: The virtual visual servoing framework. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):615–628.

[Daft and Lengel, 1986] Daft, R. L. and Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, 32(5):554–571.

[Danev et al., 2017] Danev, L., Hamann, M., Fricke, N., Hollarek, T., and Paillacho, D. (2017). Development of animated facial expressions to express emotions in a robot: Roboticon. In *2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6.

[Davis et al., 2003] Davis, L., Rolland, J., Hamza-Lup, F., Ha, Y., Norfleet, J., and Imielinska, C. (2003). Enabling a continuum of virtual environment experiences. *IEEE Computer Graphics and Applications*, 23(2):10–12.

[Debevec, 1998] Debevec, P. (1998). Rendering with natural light. In *ACM SIGGRAPH 98 Electronic Art and Animation Catalog*, SIGGRAPH '98, pages 166–, New York, NY, USA. ACM.

[Demir and Karaarslan, 2018] Demir, O. F. and Karaarslan, E. (2018). Augmented reality application for smart tourism: Gokovar. In *2018 6th International Istanbul Smart Grids and Cities Congress and Fair (ICSG)*, pages 164–167.

[Emerson, 2015] Emerson, S. (2015). Visual Effects at LAIKA, A Crossroads of Art and Technology. *ACM SIGGRAPH 2015 Talks*, page 2799649.

[Everitt et al., 2003] Everitt, K. M., Klemmer, S. R., Lee, R., and Landay, J. A. (2003). Two worlds apart: Bridging the gap between physical and virtual media for distributed design collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 553–560, New York, NY, USA. ACM.

[Fairchild et al., 2017] Fairchild, A. J., Campion, S. P., García, A. S., Wolff, R., Fernando, T., and Roberts, D. J. (2017). A mixed reality telepresence system for collaborative space operation. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):814–827.

[Feiner et al., 1993] Feiner, S., Macintyre, B., and Seligmann, D. (1993). Knowledge-based augmented reality. *Communications of the ACM*, 36(7):53–62.

[Fernando, 2005] Fernando, R. (2005). Percentage-closer soft shadows. In *ACM SIGGRAPH 2005 Sketches on - SIGGRAPH '05*, page 35.

[Franke and Jung, 2008] Franke, T. and Jung, Y. (2008). Precomputed radiance transfer for X3D based mixed reality applications. *Proceedings of the 13th international symposium on 3D web technology - Web3D '08*, (3):7.

[Gabor, 1948] Gabor, D. (1948). Holography, 1948—1971. *Proceedings of the IEEE*, 60(6):655–668.

[Giambattista della Porta, 1584] Giambattista della Porta (1584). Magia Naturalis. *Popular Science*.

[Goferman et al., 2010] Goferman, S., Zelnik-Manor, L., and Tal, A. (2010). Context-aware saliency detection. *IEEE Conference on Computer Vision and Pattern Recognition*.

[Goldstein, 1994] Goldstein, J. H. (1994). *Toys, Play, and Child Development*. Cambridge University Press.

[Grosch, 2005] Grosch, T. (2005). Differential Photon Mapping-Consistent Augmentation of Photographs with Correction of all Light Paths. *Eurographics*, pages 53–56.

[Grosch et al., 2007] Grosch, T., Eble, T., and Mueller, S. (2007). Consistent interactive augmentation of live camera images with correct near-field illumination. *Proceedings of the 2007 ACM symposium on Virtual reality software and technology - VRST '07*, 1(212):125.

[Haines and Haines, 1992] Haines, K. and Haines, D. (1992). Computer Graphics for Holography. *IEEE Computer Graphics and Applications*, 12(1):37–46.

[Haller et al., 2003] Haller, M., Drab, S., and Hartmann, W. (2003). A real-time shadow approach for an augmented reality application using shadow volumes. *Proceedings of the ACM symposium on Virtual reality software and technology - VRST '03*.

[Herling and Broll, 2012] Herling, J. and Broll, W. (2012). PixMix: A real-time approach to high-quality Diminished Reality. In *ISMAR 2012 - 11th IEEE International Symposium on Mixed and Augmented Reality 2012, Science and Technology Papers*, pages 141–150.

[Ishii et al., 2002] Ishii, H., Underkoffler, J., Chak, D., Piper, B., Ben-Joseph, E., Yeung, L., and Kanji, Z. (2002). Augmented Overlaying Drawings , Urban Planning Workbench : Physical Models and Digital Simulation. *Proceedings of the International Symposium on Mixed and Augmented Reality*, pages 1–9.

[Iwai et al., 2014] Iwai, D., Nagase, M., and Sato, K. (2014). Shadow removal of projected imagery by occluder shape measurement in a multiple overlapping projection system. *Virtual Real.*, 18(4):245–254.

[Jacobs et al., 2005] Jacobs, K., Nahmias, J.-D., Angus, C., Reche, A., Loscos, C., and Steed, A. (2005). Automatic generation of consistent shadows for augmented reality. *ACM International Conference Proceeding Series; Vol. 112*, page 113.

[Jantke et al., 2013] Jantke, K. P., Arnold, O., and Spundflasch, S. (2013). Aliens on the bus: A family of pervasive games. In *2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE)*, pages 387–391.

[Jarusirisawad and Saito, 2007] Jarusirisawad, S. and Saito, H. (2007). Diminished reality via multiple hand-held cameras. *2007 1st ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC*, pages 251–258.

[Jaynes et al., 2001] Jaynes, C., Webb, S., Steele, R. M., Brown, M., and Seales, W. B. (2001). Dynamic shadow removal from front projection displays. In *Proceedings Visualization, 2001. VIS '01.*, pages 175–555.

[Jung et al., 2007] Jung, Y., Franke, T., Dähne, P., and Behr, J. (2007). Enhancing X3D for advanced MR appliances. *Proceedings of the twelfth international conference on 3D web technology - Web3D '07*, page 27.

[Kajiya, 1986] Kajiya, J. T. (1986). The rendering equation. *SIGGRAPH Comput. Graph.*, 20:143–150.

[Kakuta et al., 2008] Kakuta, T., Vinh, L. B., Kawakami, R., Oishi, T., and Ikeuchi, K. (2008). Detection of moving objects and cast shadows using a spherical vision camera for outdoor mixed reality. In *Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology*, VRST '08, pages 219–222, New York, NY, USA. ACM.

[Karl et al., 2010] Karl, B., Staffan, J., and Staffan, B. (2010). Undercurrents: A computer-based gameplay tool to support tabletop roleplaying. In *DiGRA Nordic: Proceedings of the 2010 International DiGRA Nordic Conference: Experiencing Games: Games, Play, and Players*.

[Karsch et al., 2011] Karsch, K., Hedau, V., Forsyth, D., and Hoiem, D. (2011). Rendering synthetic objects into legacy photographs. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, SA '11, pages 157:1–157:12, New York, NY, USA. ACM.

[Kasapakis et al., 2013] Kasapakis, V., Gavalas, D., and Bubaris, N. (2013). Pervasive games research: A design aspects-based state of the art report. In *Proceedings of the 17th Panhellenic Conference on Informatics*, pages 152–157.

[Kawai et al., 2013] Kawai, N., Sato, T., and Yokoya, N. (2013). Diminished reality considering background structures. In *2013 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2013*, pages 259–260.

[Kennedy et al., 1996] Kennedy, J., Mitchell, K., and Barclay, P. (1996). A framework for information visualisation. *SIGMOD Record*, 25(4):30–34.

[Kiyokawa et al., 2002] Kiyokawa, K., Billinghurst, M., Hayes, S. E., Gupta, A., Sannohe, Y., and Kato, H. (2002). Communication behaviors of co-located users in collaborative ar interfaces. In *Proceedings. International Symposium on Mixed and Augmented Reality*, pages 139–148.

[Kjeldskov et al., 2009] Kjeldskov, J., Paay, J., O'Hara, K., Smith, R., and Thomas, B. (2009). Frostwall: A dual-sided situated display for informal collaboration in the corridor. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, OZCHI '09, pages 369–372, New York, NY, USA. ACM.

[Knecht et al., 2011] Knecht, M., Dünser, A., Traxler, C., Wimmer, M., and Grasset, R. (2011). A Framework For Perceptual Studies In Photorealistic Augmented Reality. *Proceedings of the 3rd IEEE VR 2011 Workshop on Perceptual Illusions in Virtual Environments*, pages 27–32.

[Knecht et al., 2010] Knecht, M., Traxler, C., Mattausch, O., Purgathofer, W., and Wimmer, M. (2010). Differential instant radiosity for mixed reality. In *Proceedings of the 2010 IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2010)*, pages 99–107. Best Paper Award!

[Knorr and Kurz, 2014] Knorr, S. B. and Kurz, D. (2014). Real-time illumination estimation from faces for coherent rendering. *ISMAR 2014 - IEEE International Symposium on Mixed and Augmented Reality - Science and Technology 2014, Proceedings*, (September):349–350.

[Kooper and MacIntyre, 2003] Kooper, R. and MacIntyre, B. (2003). Browsing the real-world wide web: Maintaining awareness of virtual information in an ar information space. *International Journal of Human Computer Interaction*, 16(3).

[L. Frank Baum, 1901] L. Frank Baum (1901). The Master Key: an Electrical Fairy Tale. *Bowen-Merrill*, page 102.

[Leão et al., 2011] Leão, C. W. M., Lima, J. P., Teichrieb, V., Albuquerque, E. S., and Keiner, J. (2011). Altered reality: Augmenting and diminishing reality in real time. In *Proceedings - IEEE Virtual Reality*, pages 219–220.

[Ledermann and Schmalstieg, 2005] Ledermann, F. and Schmalstieg, D. (2005). APRIL: a high-level framework for creating augmented reality presentations. *IEEE Proceedings. VR 2005. Virtual Reality, 2005.*, 2005:187–194.

[Lee et al., 2017] Lee, Y. Y., Ahmed, B., Lee, J. H., An, H., and Lee, K. H. (2017). Augmenting Three-dimensional Effects in Digital Exhibition of a Cultural Artifact using 3D Pseudo Hologram. *Adjunct Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct 2016*, pages 284–287.

[Lehtonen et al., 2016] Lehtonen, M. F., Arvo, J., and T. (2016). Extended panorama tracking algorithm for augmenting virtual 3D objects in outdoor environments. *2016 22nd International Conference on Virtual System Multimedia (VSMM)*, pages 1–8.

[Liao et al., 2010] Liao, H., Inomata, T., Sakuma, I., and Dohi, T. (2010). 3-D augmented reality for MRI-guided surgery using integral videography autostereoscopic image overlay. *IEEE Transactions on Biomedical Engineering*, 57(6):1476–1486.

[Lindley, 2005] Lindley, C. A. (2005). Game space design foundations for trans-reality games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, pages 397–404.

[Magnenat et al., 2015] Magnenat, S., Ngo, D. T., Zünd, F., Ryffel, M., Noris, G., Rothlin, G., Marra, A., Nitti, M., Fua, P., Gross, M., and Sumner, R. W. (2015). Live Texturing of Augmented Reality Characters from Colored Drawings. *IEEE Transactions on Visualization and Computer Graphics*, 21:1201–1210.

[Mandl et al., 2017] Mandl, D., Yi, K. M., Mohr, P., Roth, P. M., Fua, P., Lepetit, V., Schmalstieg, D., and Kalkofen, D. (2017). Learning lightprobes for mixed reality illumination. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 82–89.

[Mann, 2001] Mann, S. (2001). Fundamental issues in mediated reality, WearComp, and camera-based augmented reality. *Fundamentals of Wearable Computers and Augmented Reality, Lawrence Erlbaum Associates, Inc*, pages 295–328.

[Mann, 2002] Mann, S. (2002). Mediated Reality with implementations for everyday life. *MIT Press journal PRESENCE: Teleoperators and Virtual Environments*.

[Maziah et al., 2016] Maziah, N., Barkhaya, M., Dayana, N., and Halim, A. (2016). A review of application of 3D hologram in education: A meta-analysis. *2016 IEEE 8th International Conference on Engineering Education (ICEED)*, pages 257–260.

[Milgram et al., 1994] Milgram, P., Takemura, H., Utsumi, a., and Kishino, F. (1994). Mixed Reality (MR) Reality-Virtuality (RV) Continuum. *Systems Research*, 2351(Telemanipulator and Telepresence Technologies):282–292.

[Mine et al., 2012] Mine, M. R., Van Baar, J., Grundhöfer, A., Rose, D., and Yang, B. (2012). Projection-based augmented reality in Disney theme parks. *Computer*, 45(7):32–40.

[Mori et al., 1970] Mori, M., MacDorman, K. F., and Kageki, N. (1970). The uncanny valley. *IEEE Robotics and Automation Magazine*, 19(2):98–100.

[Mur-Artal et al., 2015] Mur-Artal, R., Montiel, J. M. M., and Tardós, J. D. (2015). Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163.

[Nassani et al., 2015] Nassani, A., Bai, H., Lee, G., and Billinghurst, M. (2015). Tag it!: Ar annotation using wearable sensors. In *SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications*, New York, NY, USA. ACM.

[Niantic and Company, 2016] Niantic and Company, T. P. (2016). Pokemon go.

[Nijholt, 2005] Nijholt, A. (2005). Meetings in the virtuality continuum: send your avatar. In *2005 International Conference on Cyberworlds (CW'05)*, pages 8 pp.–82.

[Nowrouzezahrai et al., 2011] Nowrouzezahrai, D., Geiger, S., Mitchell, K., Sumner, R., Jarosz, W., and Gross, M. (2011). Light Factorization for Mixed-Frequency Shadows in Augmented Reality. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011*.

[Noyes, 1964] Noyes, E. (1964). Clay or the origin of the species. *Carpenter Center For Visual Arts Harvard*.

[Park et al., 2016] Park, N.-Y., Kim, E., Lee, J., and Woo, W. (2016). All-in-One Mobile Outdoor Augmented Reality Framework for Cultural Heritage Sites. *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 484–489.

[Ramamoorthi and Hanrahan, 2001] Ramamoorthi, R. and Hanrahan, P. (2001). An efficient representation for irradiance environment maps. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, pages 497–500.

[Raskar et al., 1998a] Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L., and Fuchs, H. (1998a). The Office of the Future : A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. *SIGGRAPH '98 Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 1–10.

[Raskar et al., 1998b] Raskar, R., Welch, G., and Fuchs, H. (1998b). Spatially Augmented Reality. *IWAR '98: Proceedings of the International Workshop on Augmented Reality*, (919):63–72.

[Raskar et al., 2001] Raskar, R., Welch, G., Low, K., and Bandyopadhyay, D. (2001). Shader Lamps: Animating Real Objects with Image-Based Illumination. *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pages 89–102.

[Renevier and Nigay, 2001] Renevier, P. and Nigay, L. (2001). Mobile collaborative augmented reality: The augmented stroll. In *Proceedings of the 8th IFIP International Conference on Engineering for Human-Computer Interaction*, EHCI '01, pages 299–316, London, UK, UK. Springer-Verlag.

[Rheiner, 2014] Rheiner, M. (2014). Birdly an attempt to fly. In *ACM SIGGRAPH 2014 Emerging Technologies*, pages 3:1–3:1.

[Rosenberg, 1992] Rosenberg, L. B. (1992). The Use of Virtual Fixtures as Perceptual Overlays to Enhance Operator Performance in Remote Environments. *Armstrong Laboratory*.

[Sato et al., 1999] Sato, I., Sato, Y., and Ikeuchi, K. (1999). Illumination distribution from shadows. *Computer Vision and Pattern Recognition (CVPR)*, 1(3):306–312.

[Sekiguchi et al., 2004] Sekiguchi, D., Inami, M., Kawakami, N., and Tachi, S. (2004). The design of internet-based robotphone.

[Sekiguchi et al., 2001] Sekiguchi, D., Inami, M., and Tachi, S. (2001). Robotphone: Rui for interpersonal communication. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 277–278.

[Setlur et al., 2005] Setlur, V., Takagi, S., Raskar, R., Gleicher, M., and Gooch, B. (2005). Automatic image retargeting. *ACM International Conference on Mobile and Ubiquitous Multimedia*.

[Shannon and Weaver, 1949] Shannon, C. and Weaver, W. (1949). The mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

[Shelton and Hedley, 2002] Shelton, B. E. and Hedley, N. R. (2002). Using Augmented Reality for Teaching Earth-Sun Relationships to undergraduate geography students. *ART 2002 - 1st IEEE International Augmented Reality Toolkit Workshop, Proceedings*, pages 1–8.

[Sodnik et al., 2006] Sodnik, J., Tomazic, S., Grasset, R., Duenser, A., and Billinghurst, M. (2006). Spatial sound localization in an augmented reality environment. In *Proceedings of the 18th Australia Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, OZCHI '06, New York, NY, USA. ACM.

[Steed et al., 2012] Steed, A., Steptoe, W., Oyekoya, W., Pece, F., Weyrich, T., Kautz, J., Friedman, D., Peer, A., Solazzi, M., Tecchia, F., Bergamasco, M., and Slater, M. (2012). Beaming: An asymmetric telepresence system. *IEEE Computer Graphics and Applications*, 32(6):10–17.

[Sutherland, 1968] Sutherland, I. E. (1968). A head-mounted three dimensional display. In *Proceedings of the AFIPS '68 (Fall, part I)*, pages 757–764.

[Takeuchi and Perlin, 2012] Takeuchi, Y. and Perlin, K. (2012). ClayVision: the (elastic) image of the city. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2411–2420.

[Tang et al., 2012] Tang, J., Marlow, J., Hoff, A., Roseway, A., Inkpen, K., Zhao, C., and Cao, X. (2012). Time travel proxy: Using lightweight video recordings to create asynchronous, interactive meetings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA. ACM.

[Unity, 2019] Unity (2019). Unity Game Engine. https://unity.com/.

[Vuforia, 2019] Vuforia (2019). Vuforia Object Recognition. https://library.vuforia.com/.

[Wagner and Schmalstieg, 2003] Wagner, D. and Schmalstieg, D. (2003). First steps towards handheld augmented reality. *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pages 127–135.

[Wahab et al., 2016] Wahab, N., Hasbullah, N., Ramli, S., and Zainuddin, N. Z. (2016). Verification of a Battlefield Visualization Framework in Military Decision Making using Holograms (3D) and multi-touch technology. *2016 International Conference on Information and Communication Technology (ICICTM)*, (May):23–26.

[Weber et al., 2018] Weber, H., Prevost, D., and Lalonde, J.-F. (2018). Learning to estimate indoor lighting from 3d objects. *Computer Vision and Pattern Recognition*.

[Yang et al., 2017] Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., and Li, H. (2017). High-resolution image inpainting using multi-scale neural patch synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4076–4084.

[Yang et al., 2011] Yang, L., Tse, Y.-C., Sander, P. V., Lawrence, J., Nehab, D., Hoppe, H., and Wilkins, C. L. (2011). Image-based bidirectional scene reprojection. *ACM Trans. Graph.*, 30(6):150:1–150:10.

[Yeh et al., 2017] Yeh, R. A., Chen, C., Lim, T. Y., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. (2017). Semantic image inpainting with deep generative models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6882–6890.

[Yim and Shaw, 2011] Yim, J.-D. and Shaw, C. D. (2011). Design considerations of expressive bidirectional telepresence robots. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA. ACM.

[Zarzycki, 2012] Zarzycki, A. (2012). Urban games: Inhabiting real and virtual cities. In *Proceedings of the 30th eCAADe Conference - Volume 1 / ISBN 978-9-4912070-2-0, Czech Technical University in Prague, Faculty of Architecture (Czech Republic) 12-14 September 2012, pp. 755-764.*

[Zhang et al., 2015] Zhang, L., Zhang, Q., and Xiao, C. (2015). Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636.

[Zheng and Wang, 2013] Zheng, X. and Wang, L. (2013). A Video-based Interface for Hand-Driven Stop Motion Animation Production. *IEEE Computer Graphics and Applications*.

[Zilly et al., 2016] Zilly, F., Ziegler, M., Keinert, J., Schoberl, M., and Foessel, S. (2016). Computational Imaging for Stop-Motion Animated Video Productions. *SMPTE Motion Imaging Journal*, 125(1):42–47.

[Zünd et al., 2015] Zünd, F., Ryffel, M., Magnenat, S., Marra, A., Nitti, M., Kapadia, M., Noris, G., Mitchell, K., Gross, M., and Sumner, R. W. (2015). Augmented creativity: Bridging the real and virtual worlds to enhance creative play. In *SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications*, SA '15, pages 21:1–21:7, New York, NY, USA. ACM.

Appendix A

Making Of - *"The Girl & the Purse"*

Fig. A.1 Mannequin made of plastic clay with interchangeable facial expressions used for the stop motion animation clip.

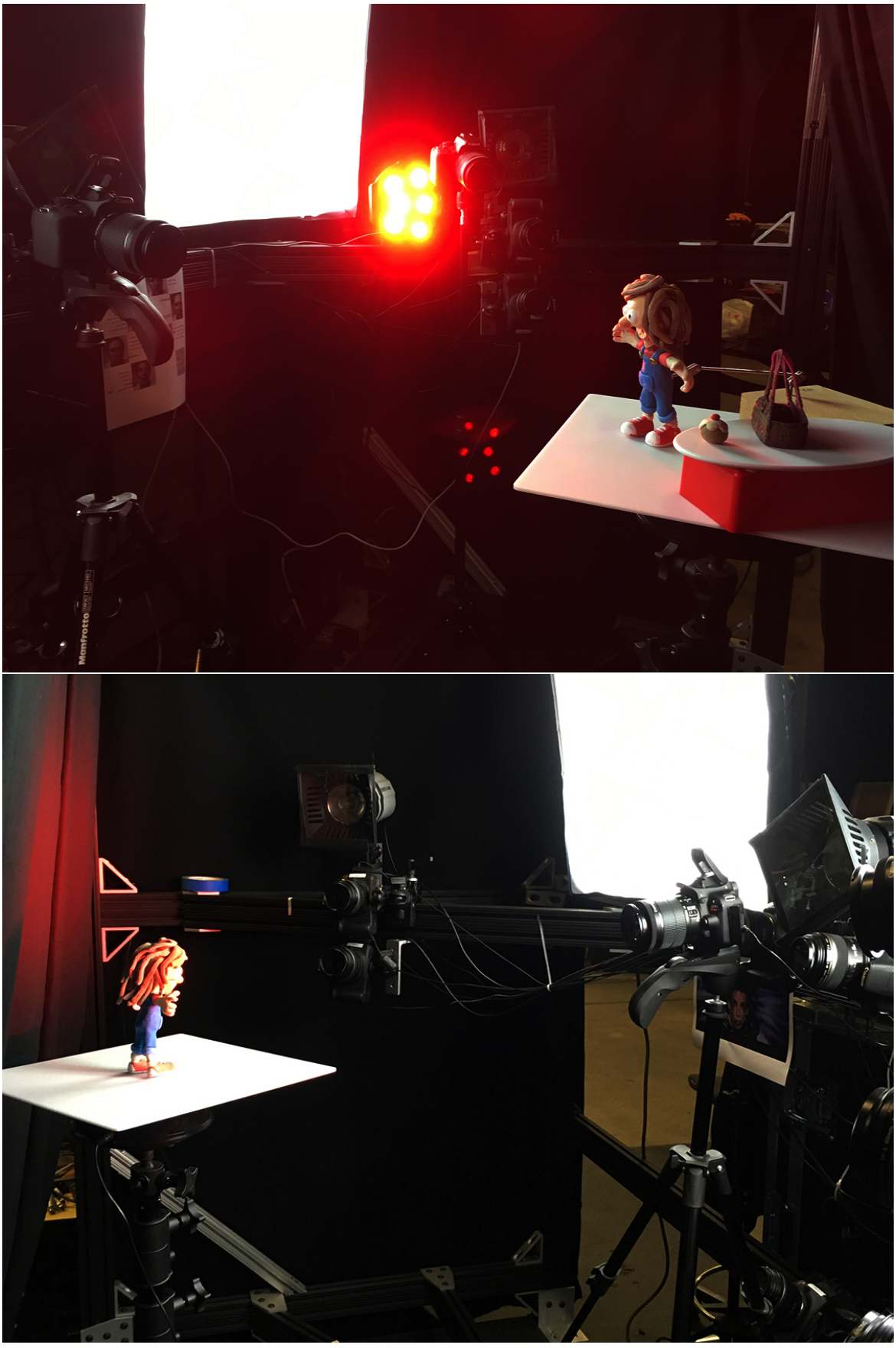


Fig. A.2 Lighting setup and camera rig used for shooting the stop motion animation clip.

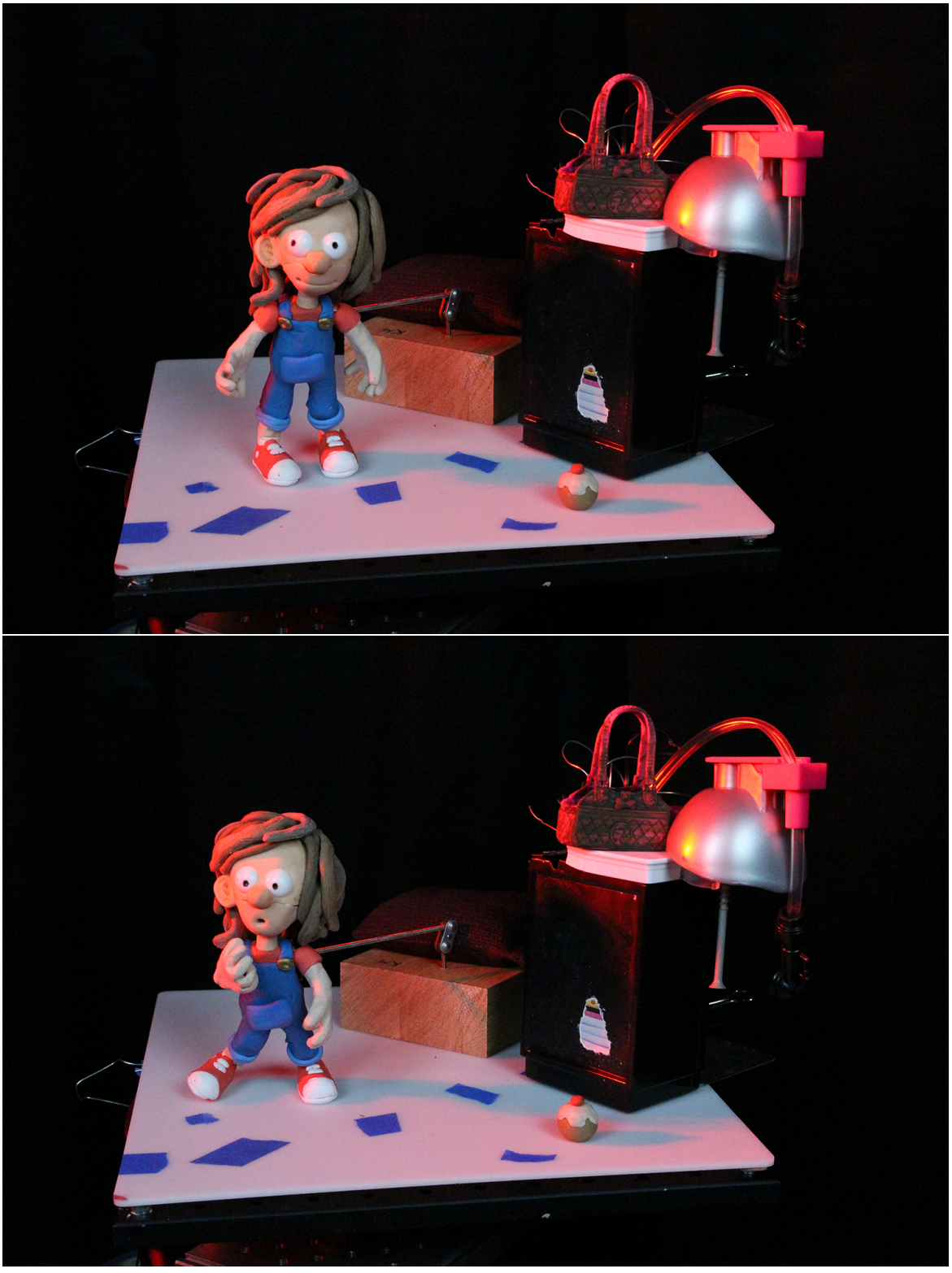


Fig. A.3 Real-world key-frame poses captured from the stop motion animation clip.

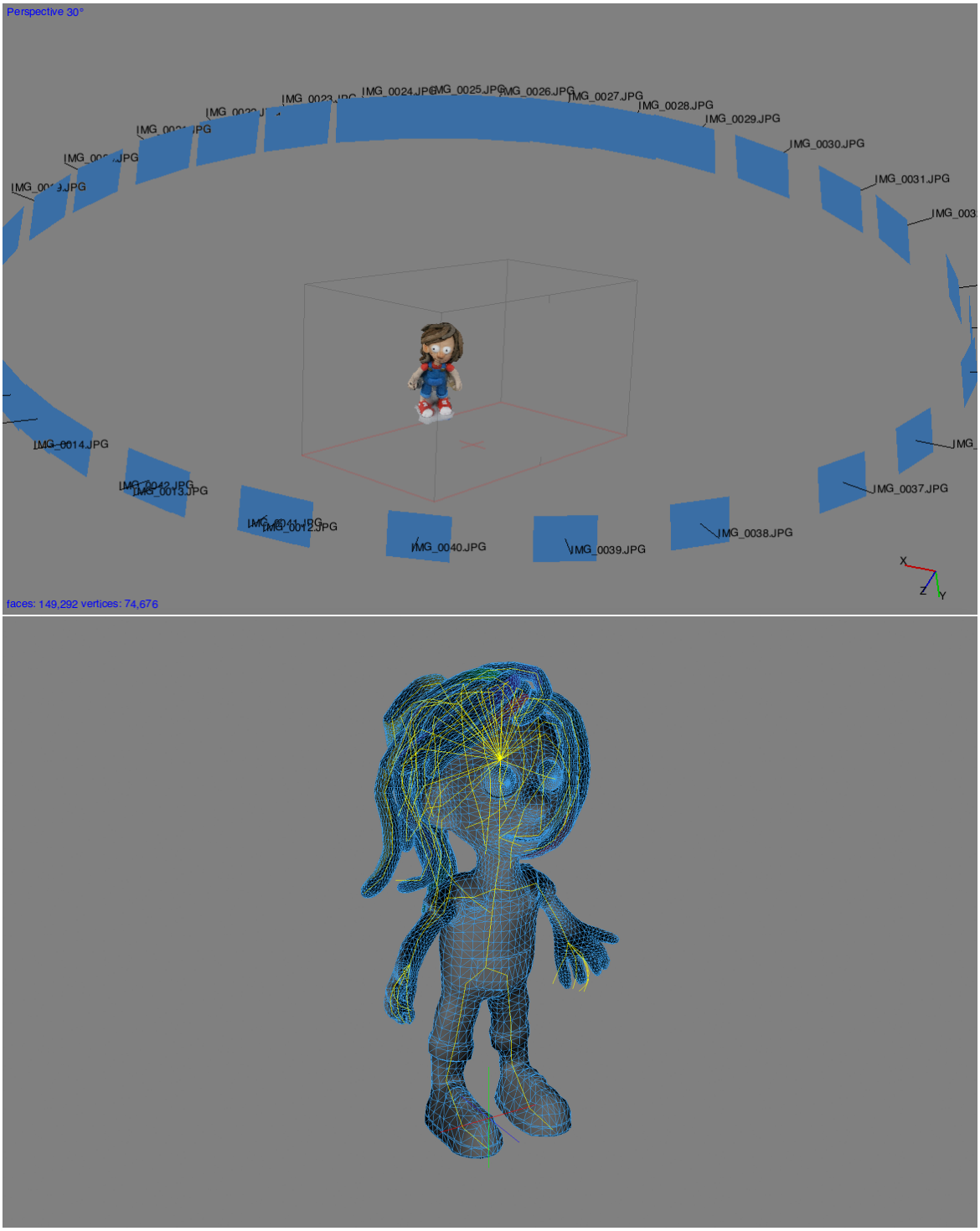


Fig. A.4 *Top)* 3D reconstruction of the real-world puppet using photogrammetric software. *Bottom)* Reconstructed and rigged puppet used for augmented stop motion animation.