Ref. Ares(2023)6848322 - 09/10/2023

Performing arts in a new era

D2.5 – Performance capturing, storing protocols and VR streaming v1

30 September 2023 – M12

Document identifier: PRMR-D2.5-Performance capturing, storing protocols and VR streaming v1
Version: 1
Author: Panayiotis Charalambous,
Dissemination status: PU
Dissemination status: PU

Grant Agreement n°	101061303
Project acronym	PREMIERE
Project title	Performing arts in a new era: AI and XR tools for better understanding, preservation, enjoyment and accessibility
Funding Scheme	HORIZON-CL2-2021-HERITAGE-01 (HORIZON Research and Innovation Actions)
Project Duration	01/10/2022 – 30/09/2025 (36 months)
Coordinator	Athena Research Center (ARC)
Associated Beneficiaries	• Stichting Amsterdamse Hogeschool voor de Kunsten (AHK) • Forum Danca - Associacao Cultural (FDA) • Tempesta Media SL (TMP) • Cyens - Centre of Excellence (CNS) • Kallitechniki Etaireia Argo (ARG) • Medidata.Net - Sistemas de Informacao para Autarquias SA (MED) • Fitei Festival Internacional Teatro Expresao Iberica Crl (FIT) • Instituto Stocos (STO) • Universite Jean Monnet Saint-Etienne (UJM) • Associacao dos Amigos do Coliseu Doporto (COL) • Stichting International Choreographic Arts Centre (ICK)

Project no. 101061303
PREMIERE

Performing arts in a new era: AI and XR tools for better understanding, preservation, enjoyment and accessibility

HORIZON-CL2-2021-HERITAGE-01

Start date of project: 01/10/2022

Duration: 36 months

History Chart				
Issue	**Date**	**Changed page(s)**	**Cause of change**	**Implemented by**
0.1		-	Structure	Panayiotis Charalambous (CNS)
0.2	14/9/2023	All	Motion Capture technologies	Panayiotis Charalambous (CNS) Mercè Álvarez (CNS)
0.3	27/9/2023	All	Streaming	Panayiotis Charalambous Mercè Álvarez (CNS)
1.0	2/10/2023	All	Internal Review	Panayiotis Charalambous (CNS)

Validation			
No.	**Action**	**Beneficiary**	**Date**
1	Prepared	Aggelos Gkiokas (ARC) Daniel Bisig (STO)	2/10/2023
2	Approved	Aggelos Gkiokas (ARC) Daniel Bisig (STO)	6/20/2023
3	Released	Aggelos Gkiokas (ARC) Daniel Bisig (STO)	9/10/2023

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Table of Contents

List of figures

List of tables

Executive Summary

In today's ever-evolving world of multimedia, the capturing and representation of human performances in digital format is not just a technological endeavor but an art form in itself. The focus of this document is to delve into the intricate world of performance capture systems for theatre and dance, offering a comprehensive analysis of their taxonomy, components, advantages, and limitations.

Starting with a thorough exploration of motion capture systems, we differentiate between two primary categories: Outside-In and Inside-Out motion capture (Section 2). While the former utilizes external cameras or sensors to capture movement from the outside, the latter leverages sensors placed on the performer. Each has its unique set of advantages, challenges, and applications, making them suited for specific use cases. They differ primarily in how they track and perceive the objects being monitored; “outside-in” systems use tracking devices that are fixed in the environment whereas “inside-out” systems use tracking devices that are fixed on the tracked subject (Figure 2.1).

Outside-in motion capture systems are characterized by using external cameras or sensors placed around the area where motion is to be tracked. These systems are commonly used in applications such as animation, biomechanical analysis, sports performance analysis, and research. Inside-out motion capture systems are characterized by having tracking sensors or cameras located on or near the object being tracked (Figure 2.1). These systems are often used in applications like human motion capture, virtual reality (VR), augmented reality (AR), and robotics. They are typically cheaper than outside-in systems, with less accuracy and are more portable. Here are some key details and considerations for inside-out motion capture systems.

Beyond motion, capturing other modalities such as audio, lighting, and props is crucial for a holistic representation of performances (Section 3). The document provides insights into the techniques and technologies associated with capturing these modalities, ensuring that every nuance of a performance is digitally preserved. Capturing accurate lighting is vital not only for the visual integrity of a performance but also for understanding and recreating the ambiance and mood of a scene. It plays a pivotal role in the portrayal of realism, depth, and aesthetics. Capturing the movement and interaction of props adds another layer of complexity and detail to a performance. Whether it's a sword in a combat scene or a ball in a dance sequence, props can significantly influence the narrative and emotion of a performance. A very important prop in many productions is clothing which is a non-rigid body, and it is quite challenging to capture. Here's what to consider in the capturing of props:

Selection of the right capture system is paramount. Thus, we discuss the fundamental criteria to consider (Section 4). Given the varied needs of different artistic endeavors, distinct requirements for dance and theater captures are also presented. Budget considerations are foundational when selecting a set of performance capture systems. The financial resources available for the project significantly influence the options and capabilities of the chosen systems. The level of intrusiveness associated with a performance capture system is a crucial consideration. Some systems may require performers to wear sensors, markers, or specialized attire, impacting their freedom of movement and comfort. Evaluating the level of intrusiveness ensures that the chosen system aligns with the performers' needs and the project's objectives. The scale of a performance, including the number of performers involved, influences the choice of a capture system. Some systems are optimized for single performers or small groups, while others excel in capturing large ensembles. Determining the number of performers is vital in selecting a system that can effectively handle the intended

scope of the project. Performance capture can focus on capturing pose (the orientation and configuration of body parts), positions (the spatial location of objects or performers) or both. This criterion is vital for projects where precision in either pose or position data is essential. Facial expressions and emotions are integral aspects of many performances. Assessing a performance capture system's ability to capture facial movements and expressions is critical for projects where facial fidelity is essential, such as character-driven animations for movies or theatre or clinical assessments. Certain applications demand high-fidelity hand and finger capture, while others may prioritize broader body movements.

In the context of the PREMIERE project, we analyze four distinct use cases, each with its unique set of requirements (Section 5). These range from AR/VR-enhanced archive browsing to real-time VR-enhanced live performances and virtual co-creation spaces. The document offers insights into the specific needs of each use case, helping stakeholders make informed decisions.

Performance capture in PREMIERE involves a) data collection of performers and their environment from the different physical locations, b) local synchronization of these data, c) transmission of data in standardized formats through a network to the virtual theatre state synchronization server which then d) ensures all performers experience the same state and e) storage of appropriate data into the CMS. In the following paragraphs we analyze how data should be captured from each local performance space. In Section 6, which deals with performance capture protocols delves into the intricacies of capture devices and the types of output they produce. We also touch upon the significance of audio in both waveform and MIDI formats.

Synchronization is a fundamental element in networked performances, ensuring that participants can collaborate effectively, deliver their artistic expressions with precision, and provide audiences with a cohesive and engaging experience. Whether it's a live-streamed dance performance, a virtual theater production, or a collaborative online art project, synchronization plays a vital role in the success and impact of networked performances. PREMIERE should be synchronized at different levels; a) all devices on a physical location should be synchronized through a local network using timecode and potentially genlock, b) the virtual theatre should be perceived synchronously by performers at a global network level (i.e., the internet) and finally different users should perceive/receive a synchronized view of the performance and any relevant data. The crucial aspect of synchronization of devices and performers state as well as the potential of VR streaming, particularly through technologies like CloudXR, is explored in Section 7. As the realms of VR and performance art converge, such technologies promise to revolutionize how audiences experience performances.

In conclusion, this document serves as a comprehensive guide to the world of performance capture, offering insights, clarifications, and direction to artists, technologists, and enthusiasts alike.

Acronyms and abbreviations

Acronym	Abbreviation
MOCAP	Motion Capture
VR	Virtual Reality
AR	Augmented Reality
MIDI	Musical Instrument Digital Interface
OSC	Open Sound Control
XR	eXtended Reality
BVH	Biovision Hierarchy
EOS	Epic Online Services
HDR	High Dynamic Range
SDR	Standard Dynamic Range
LAN	Local Area Network
RFC	Remote Function Calls

1. Introduction

Motion capture systems aim to capture and track the movement of objects or, more commonly, human bodies for various applications. Motion capture techniques have been widely used in the entertainment industry for many years, especially in the production of animated movies, video games, and virtual reality experiences. With the advent of new hardware and software technologies, such as high-speed cameras, inertial measurement units, and depth sensors, motion capture has become more sophisticated, precise and even affordable, allowing for the capture of even the most subtle movements and expressions.

Performance capture takes this a step further by capturing not only the body movements but also the facial expressions, eye movements, and even the voice of actors. This technique allows for a more realistic and nuanced portrayal of characters in digital media, and has been used in movies like Avatar[1], Planet of the Apes[2], and The Lord of the Rings series of movies[3].

In addition to entertainment, motion and performance capture have also found applications in fields such as sports training, biomechanics research, and medical rehabilitation. By analysing the movements of athletes or patients, for example, researchers can gain insights into how to improve performance or aid in recovery from injuries.

Overall, the increasing sophistication of motion and performance capture technologies is opening up new possibilities for how we can represent and interact with the digital world and is likely to continue to play a significant role in the entertainment and technology industries in the years to come.

In this document, our objective is to explore potential technologies that the PREMIERE consortium can leverage to fulfil various use cases based on the first version of the PREMIERE architecture; an overview of that system is shown in Figure 1.1. Performance capture touches two components of the PREMIERE architecture; the local capture of a performer and everything related to the performance and the network synchronization of performers in different physical locations in a virtual space (highlighted in red). We also provide guidelines for capturing and streaming live and recorded performances. Ultimately, this document serves as a valuable resource for those interested in the digital transformation of the performing arts.

[1] Avatar Performance Capture: https://youtu.be/OJ1JzYPjcj0?si=eWT37pRhLtZVCzgN
[2] Terry Notary: Apes Movement Demo: https://youtu.be/3fJhZSwoCVQ?si=r8x2DPjXBrKL2ICZ
[3] The Hobbit: An Unexpected Journey VFX: https://youtu.be/HUGdsjwTf2s?si=ankxIqFmHHRdqWBE

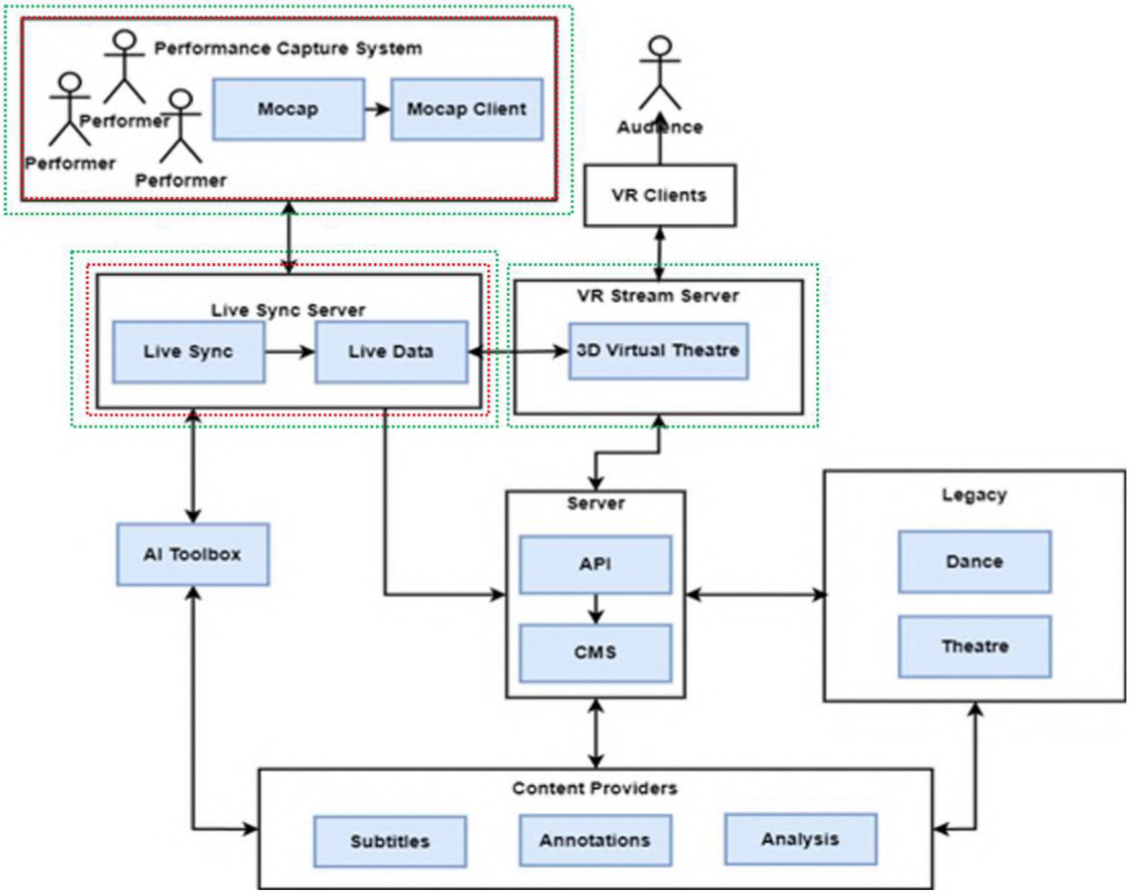

Figure 1.1 (Red) Performance capture touches two components in version 1 of the PREMIERE architecture; the local capture of a performer and the network synchronization of performers in different physical locations in a virtual space. (Green) In order to stream the data to the different end users, data need to be synchronized in three places; a) data from physical devices locally at each performing stage, b) data from each performer on a live sync server so that everyone experiences the same state and c) and synchronization of experiences for remote viewers.

2. Motion Capture Systems: A Taxonomy

Motion capture systems aim to capture and track the movement of objects or, more commonly, human bodies for various applications, including animation, gaming, biomechanical analysis, and virtual reality. Based on the way motion is being captured, these systems can broadly be categorized in “inside-out” and “outside-in” systems. They differ primarily in how they track and perceive the objects being monitored; “outside-in” systems use tracking devices that are fixed in the environment whereas “inside-out” systems use tracking devices that are fixed on the tracked subject (Figure 2.1).

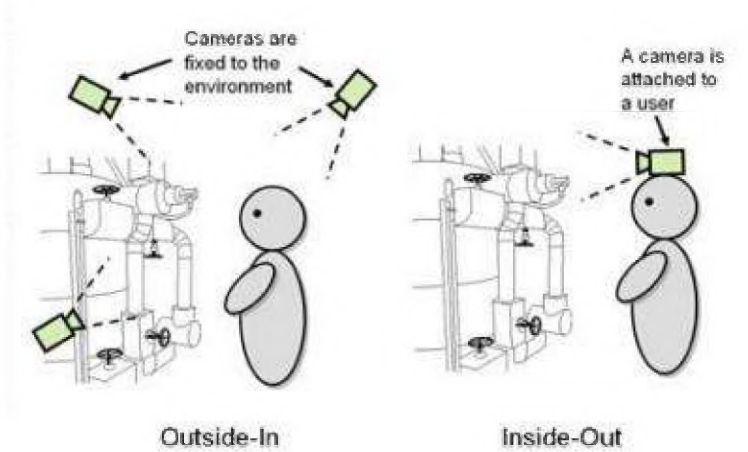

Figure 2.1: Outside-In vs Inside-out tracking[4]

2.1. Outside-In Motion Capture

Outside-in motion capture systems are characterized by using external cameras or sensors placed around the area where motion is to be tracked. These systems are commonly used in applications such as animation, biomechanical analysis, sports performance analysis, and research. Here are more details and considerations regarding outside-in motion capture systems:

2.1.1. Components

- <u>*Cameras or Sensors*</u>*:* Outside-in systems typically use multiple high-speed cameras or sensors strategically placed around the tracking area (Figure 2.2). The number of cameras can vary depending on the desired level of accuracy and tracking volume.
- <u>*Markers or Reflective Objects*</u>*:* Objects or markers with known characteristics (e.g., reflective spheres or markers with unique patterns) are typically attached to the objects or subjects being tracked. These markers are used as reference points for tracking and allow for very precise tracking. Marker can either be *passive* (reflective) or *active* (emitting); active are typically more accurate since each marker has a specific ID, however they require external power sources. There are also marker-less solutions that rely on computer vision; these are typically more flexible solutions since they do not require a fixed setup, uniforms and markers, however they are less accurate than marker-based solutions and they are computationally more expensive.

[4] Images from: Ishii, K. (2010). Augmented Reality: Fundamentals and nuclear related applications. Nuclear Safety and Simulation, 1(1)

- *Processing Unit:* Data from the cameras or sensors are collected and processed by a central unit or software to calculate the positions of the markers in three-dimensional space. Some systems allow for the definition of rigid bodies and hierarchies to allow for streaming of more structured data, such as skeletal hierarchies or rigid bodies.

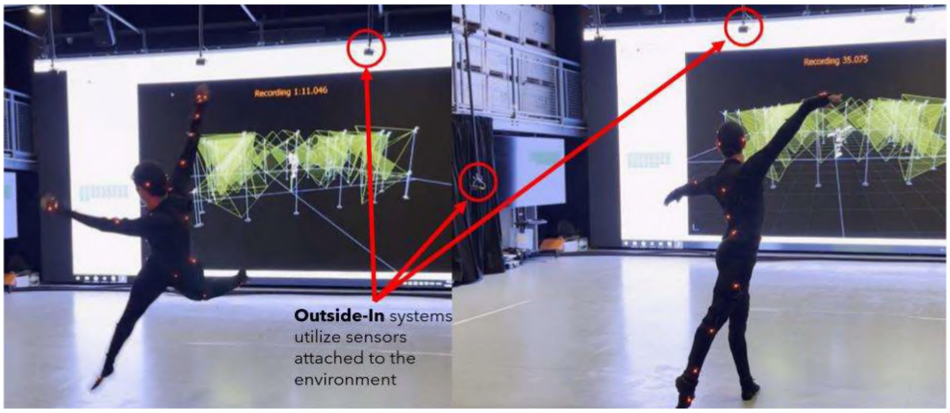

Figure 2.2: Outside-In systems have external cameras or sensors around the capture area.

2.1.2. Working Principle

- Outside-in systems track the movement of markers on the objects or subjects from the perspective of external cameras or sensors.
- Cameras continuously capture images of the markers from different angles and perspectives.
- Specialized algorithms analyse the marker positions in these images to calculate the object's position and orientation in a global-coordinate system in real-time.

2.1.3. Advantages

- Large Tracking Volume: Outside-in systems are capable of covering large tracking volumes, making them suitable for applications where objects or subjects need to move over a significant area.
- Reduced Occlusion in marker-based systems: Since external cameras (for camera-based solutions) have a wide view of the entire tracking area, occlusion issues (where markers are blocked from view) are minimized.
- High Accuracy: These systems can provide high levels of tracking accuracy, especially when using multiple cameras and sophisticated algorithms.

2.1.4. Challenges and Limitations

- Fixed Setup: Outside-in systems require a fixed setup with cameras or sensors placed in predetermined positions. This setup can be less portable and flexible compared to inside-out systems.
- *Occlusions are still an issue* when having multiple performers with lots of interactions at the same time. This is especially true in the case of dance performances with lots of contacts between dancers or the ground (e.g., when lying down). This can be helped with (but not eliminated) with the definition of rigid bodies (groups of sensors).

- Cost: Setting up and maintaining an outside-in motion capture system can be expensive due to the need for multiple high-quality cameras and specialized software.
- Marker Attachment: Attaching markers to objects or subjects can be time-consuming, especially when tracking complex movements involving many body parts or objects. Additionally, this can be problematic in situations where the subjects/actors might need to wear specific clothing such as dresses or traditional uniforms.

2.1.5. Applications

- Animation: Outside-in motion capture is widely used in the entertainment industry for capturing the movements of actors and objects to create realistic animations for movies, video games, and virtual productions.
- Biomechanics and Sports Analysis: Due to their accuracy, researchers and sports professionals use outside-in systems to analyze the movements of athletes and study human biomechanics for injury prevention and performance enhancement.
- Engineering and Product Design: Outside-in motion capture is employed in engineering and product design to study the ergonomics of products and human interactions with machines.

2.1.6. Systems for Outside-in motion capture

In this section we summarize some outside-in motion capture systems grouped by technology.

1. Optical Marker-Based Systems:

    - Vicon: Offers marker-based optical motion capture systems with a wide range of cameras and marker options. Recently, they announced that they are working on a marker-less based solution.
    - OptiTrack: Known for its marker-based motion capture technology, providing cameras and markers for high-precision tracking.
    - Qualisys: Provides optical marker-based motion capture systems suitable for biomechanics, sports science, and animation. They also offer underwater motion capture solutions.
    - PhaseSpace: Offers marker-based optical motion capture systems with a focus on low latency and high accuracy. They mostly use active markers where each marker has a specific id. These require power sources that can be an issue for specific scenarios.
    - Motion Analysis: Specializes in marker-based optical motion capture systems for applications like biomechanics and animation.
    - Synertial: Provides full-body marker-based motion capture suits for animation, virtual reality, and research.

2. Structured Light Systems:

    - Microsoft Kinect: The first versions of this system used both structured light technology for depth sensing and full-body tracking in gaming and interactive applications. This was later replaced with a Time-of-flight approach.
    - Intel RealSense Camera: projects a series of infrared patterns to obtain 3D structure. It also incorporates depth-sensing technology for facial and hand tracking, as well as 3D scanning.

3. **Time-of-flight Systems:**
   - Microsoft Azure Kinect: uses time-of-flight to measure distances and estimate depth. The first version of the Kinect used structured light.
   - Smartphone cameras: Some mobile devices from companies like Sony and LG have time-of-flight cameras.
4. **Camera-Based Systems (No Markers):**
   - Apple ARKit: Uses the built-in cameras of iOS devices for markerless motion capture and augmented reality applications.
   - Leap Motion Controller: Utilizes cameras and infrared sensors for hand and finger tracking in gesture recognition.
   - 4DViews: Uses a set of synchronized cameras to reconstruct human motion. It can also capture a volumetric representation, i.e, a mesh per frame, which essentially encodes motion and geometry of the subjects being captured.
   - Move.AI: It can track human movement using cameras from mobile devices. It can work using one or more cameras.
   - Monocular motion capture: There is a lot of research currently on how to export human motion from a single camera feed. These are typically 2D projections of motion or 3D estimates under some assumptions that are currently of low accuracy as compared to all other systems. An example is Rokoko Vision.
5. **High-Speed Cameras for Sports Analysis:**
   - Dartfish: Offers high-speed camera systems for sports analysis, enabling precise tracking of athletes' movements.
   - Raptor Photonics (Raptor-12S and Raptor-4S): Provides high-speed cameras for motion capture applications, including sports analysis.
6. **Electromagnetic Tracking Solutions:**
   - Northern Digital Inc. (NDI): Uses both Electromagnetic (e.g., Aurora) and Optical tracking technologies (e.g., Polaris) to accurately localize and track medical instruments in 3D space. Mostly used in medical research, industrial measurement, and animation.
7. **VR Headsets with Outside-In tracking**
   - HTC Vive, Valve Index: these are VR systems that utilize base stations placed around the environment to have precise tracking of sensors attached to headsets or other devices.

2.2. Inside-Out Motion Capture

Inside-out motion capture systems are characterized by having tracking sensors or cameras located on or near the object being tracked (Figure 2.1). These systems are often used in applications like human motion capture, virtual reality (VR), augmented reality (AR), and robotics. They are typically cheaper than outside-in systems, with less accuracy and are more portable. Here are some key details and considerations for inside-out motion capture systems.

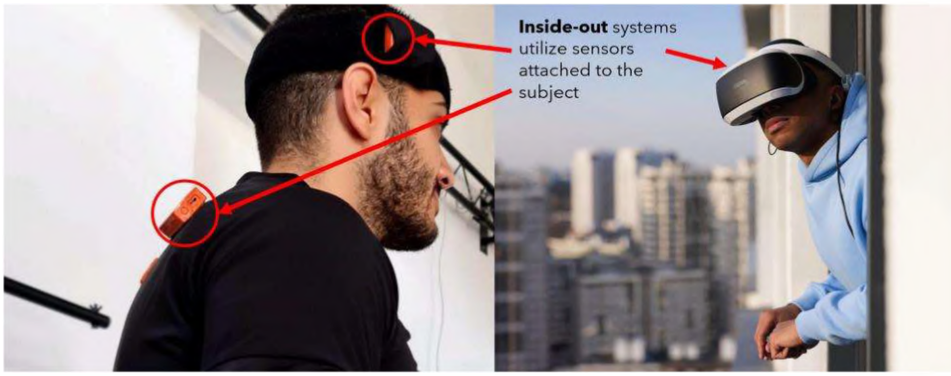

Figure 2.3 Inside-Out systems utilize sensors attached to the subjects to estimate motion[5].

2.2.1. Components

- Sensors or Cameras: Inside-out systems typically use sensors or cameras that are attached to the object being tracked. In the case of human motion capture, these sensors may be placed on key body joints or segments, such as the wrists, elbows, shoulders, and ankles.
- Processing Unit: These systems require a processing unit that collects data from the sensors or cameras and processes it to estimate the object's position and orientation in real-time.

2.2.2. Working Principle

- Inside-out systems work by tracking the movement of the sensors or markers from the perspective of the object itself.
- Cameras or sensors continuously capture data about the position and orientation of the markers on the object.
- Algorithms process this data to reconstruct the object's movement in three-dimensional space.

2.2.3. Advantages

- Freedom of Movement: Inside-out systems offer a high degree of freedom of movement because the tracking sensors move with the object. This makes them ideal for applications where the tracked object (e.g., a person) needs to move around freely.
- Portability: These systems are often more portable and versatile compared to their outside-in counterparts, making them suitable for a variety of environments.

2.2.4. Challenges and Limitations:

- Limited Tracking Volume: Inside-out systems have a limited tracking volume, which is defined by the range of motion that the sensors or markers can cover with respect to the receiver (e.g., a Router or a base station)[6]. Movella for example has three

[5] Image on the right free to use from Pexels (https://www.pexels.com/photo/man-in-blue-hoodie-wearing-a-virtual-reality-headset-7864588/).
[6] Xsens User Manual

different systems with different radius: a) the XSens MVN has a maximum range of 50 meters indoor and 150m outdoor, b) the XSens Awinda has a range of 20m for indoor and 50 for outdoor and c) the XSens Awinda Dongle has a range of 10m indoor and 20m outdoor. This can be a limitation when tracking objects or subjects over larger areas (e.g., sports).

- Occlusion Issues: If the sensors or markers on the object are blocked from the view of other sensors or cameras (occlusion), it can lead to tracking inaccuracies. This can be an issue in indoor spaces with objects (such as an office space with walls, desks, chairs) that interfere with the signals. This is especially true for systems that communicate of the 5GHz channels of Routers (e.g., Rokoko or XSens).
- Calibration and Initialization: Setting up and calibrating inside-out systems can be a complex process, especially when dealing with multiple sensors or markers.
- Accuracy: These systems are typically less accurate than outside-in systems.
- Drifting: Due to inaccuracies in sensor estimations, there is a shift in the positional data; i.e., after some time, the object appears to have shifted by some amount. This drifting is typically different for the different technologies. Some people use a combination of outside-in (e.g., VIVE trackers and base stations[7], Rokoko Coil Pro[8]), and inside-out systems to minimize drifting.

2.2.5. Applications:

- Human Motion Capture: Inside-out systems are commonly used in the entertainment industry for capturing the movements of actors for animation and special effects in movies, video games, and VR experiences.
- VR and AR: These systems are integral to many VR and AR setups, where they track the movement of headsets, controllers, and even the user's body for immersive experiences.
- Robotics: Inside-out systems are used in robotics to track the movements of robot arms, end effectors, and other components.

2.2.6. Systems for Inside-Out Motion Tracking

Inside-out systems can be split based on the underlying technology used:

1. **Inertial Motion Capture (IMU) Systems:**
   - IMU-based motion capture systems use inertial sensors, such as accelerometers and gyroscopes, attached to the body or object being tracked. These sensors measure changes in motion and orientation.
   - Examples: Xsens MVN, Xsens Awinda, Rokoko SmartSuite Pro, Noitom Perception Neuron, Sony Mocopi.
2. **Mechanical/Exoskeleton Motion Capture Systems:**
   - Mechanical motion capture systems, also known as exoskeleton motion capture systems, directly track body joint angles by attaching sensors to a skeletal-like structure worn by performers. As the performer moves, these articulated mechanical parts measure their relative motion. Some of these systems allow for the measurements of limbs for more precise data. These systems offer real-time tracking, are relatively cost-effective, suffer no occlusion issues, and are wireless with an unlimited capture volume. They

[7] https://www.vive.com/eu/accessory/base-station/

[8] https://www.rokoko.com/products/coil-pro

typically consist of rigid structures with jointed metal or plastic rods connected by potentiometers, aligning with the body's joints. Prices for these systems range from $25,000 to $75,000, with additional costs for an external absolute positioning system. Some variants may offer limited force feedback or haptic input.
    - Examples: Metamotion Gypsy
3. **VR Headsets with Inside-Out Tracking:**
    - Virtual reality headsets often incorporate inside-out tracking for the user's head and hand controllers. Cameras or depth-sensors on the headset and controllers track their positions in real-time.
    - Example: Oculus Rift S, HTC Vive, Meta Quest 2 and 3, Pico, and Meta Quest Pro.
4. **Gesture Recognition Devices:**
    - Gesture recognition devices use cameras and sensors to track hand and finger movements for applications like sign language interpretation, gaming, and human-computer interaction.
    - Example: Leap Motion Controller.
5. **Smartphones and Tablets for Motion Capture:**
    - Some mobile devices come equipped with motion sensors that can be used for basic motion capture of the devices themselves. Apps and software can turn these devices into motion capture tools for simple applications.
    - Example: Apple ARKit for iOS devices.

2.3. Summary

In summary, the main difference lies in the placement of the tracking sensors or cameras. Overall, inside-out motion capture systems offer flexibility and mobility but may have limitations in terms of tracking volume and occlusion. They are chosen based on the specific requirements of the application and the need for freedom of movement. Outside-in motion capture systems offer a larger tracking volume, reduced occlusion, and high accuracy but may require a fixed setup and involve marker attachment. They are chosen based on the specific needs of the application, such as the size of the tracking area and the desired level of precision.

To conslude, the choice depends on specific needs; inside-out systems excel in mobility, whereas outside-in systems excel in precision and coverage.

3. Capturing Other Modalities

PREMIERE aims to modernize the performing arts; more specifically dance and theatre. These art forms do not only rely on body movement, but also require the simultaneous and synchronous acquisition of other dynamic data such as audio and lighting, to create a holistic and immersive performance capture experience.

3.1. Audio

In the realm of performance capture, audio is an integral component that complements the visual data obtained from motion capture. Capturing audio adds depth and realism to the recorded performance. Here are more details and considerations regarding audio capture:

- Audio capture devices:
  - *Microphone Array (speech, music):* This allows for the capture of audio from multiple directions, useful for spatial audio and multi-performer settings.
  - *Lavalier Microphones (speech, music):* These are small clip-on microphones that can be attached directly to a performer for isolated voice capture.
  - *MIDI Instruments (music):* For capturing music or other instrumental sounds, especially in a controlled studio environment. Typical musical instruments with MIDI output include pianos, synthesizers, electric guitars, and drums.
- Key considerations for audio capture:
  - *Synchronization:* Audio and motion data must be synchronized accurately to ensure that sound aligns seamlessly with the corresponding actions. This synchronization enhances the overall authenticity of the captured performance. This is typically done with a combination of genlock for a global clock signal with the same frequency for all captured content and a timecode for timing synchronization.
  - *Microphone Placement:* Proper placement of microphones is crucial to capture clear and high-quality audio. In some cases, lavalier microphones may be attached on performers to capture their voices without interference.
  - *Environmental Factors:* Consideration must be given to the performance environment. Factors such as background noise, echoes, and ambient sound can impact audio quality. Soundproofing or controlled acoustic environments may be necessary.
  - *Multi-Channel Audio:* For more immersive experiences, multi-channel audio capture may be employed. This allows for the recording of spatial audio, which can be particularly valuable in virtual reality (VR) and augmented reality (AR) applications.
  - *Post-Processing:* After capture, audio data often undergoes post-processing to enhance its quality and optimize it for integration with the visual elements of the performance.
- Challenges in Audio Capture:

- Audio Spill: The overlapping of different audio sources can be an issue, particularly in live performances.
    - *Audio Feedback:* Special care must be taken to prevent feedback loops in a live setting, particularly when loudspeakers are also in use.
    - *Maximisation of the dynamic range* while not exceeding it
    - *Suboptimal frequency response* when recording off axis.

- **Integration with Other Modalities:**
    - *Audio-Visual Synchronization:*
        - *Timestamps:* Both audio and visual data should be timestamped to ensure accurate synchronization during post-processing or real-time streaming. Please see Section 7 for a more detailed discussion.
        - *Clapperboard Technique:* A clapperboard or a similar audio-visual cue can be used at the beginning of each take to simplify the process of aligning audio and video tracks in post-production.
    - *Audio-Motion Capture Integration:*
        - *Sound-to-Motion Mapping:* Audio data, especially vocalizations and physical sounds like footsteps or claps, can be mapped to motion capture data to enhance the realism and expressive potential of animated characters.
        - *Spatial Audio:* In VR or AR settings, audio can be spatially mapped to the location of characters or objects as captured by motion tracking, providing a more immersive experience.
    - *Multi-Sensory Experience:*
        - *Haptic Feedback:* Audio cues can trigger haptic feedback in interactive experiences, such as a vibration in a handheld controller in sync with a particular sound.
        - *Lighting Cues:* In more advanced setups, audio can be used to trigger specific lighting states or changes, thus enhancing the mood or narrative emphasis.
    - *Real-Time Interaction:*
        - *Voice Commands:* In interactive settings, real-time audio capture can be used to trigger actions or changes in the visual or motion capture elements. This requires integration with speech recognition systems.
        - *Interactive Music:* In performances involving musical instruments, motion capture of the instrumentalists can be synchronized with the audio to recreate realistic virtual performances.
        - Artistic manipulation of a scene like Interactive Sonification of Performer's Movements / Activities, performers body as musical Instrument, real-time interaction with other elements on stage such as video and/or lights.
    - *Data Packaging for Transmission:*
        - *Unified Data Streams:* Audio, video, and other sensory data can be bundled into unified data streams for easier transmission, storage, or real-time streaming.

- *Metadata:* Metadata tagging can be used to indicate how the audio relates to other modalities, such as which microphone corresponds to which performer or which sounds are linked to specific actions or markers in the motion capture data.

The inclusion of audio capture in performance systems adds a layer of realism that can elevate the final output, making it suitable for a wide range of applications, from film and gaming to live theatrical performances.

3.2. Lighting

Capturing accurate lighting is vital not only for the visual integrity of a performance but also for understanding and recreating the ambiance and mood of a scene. It plays a pivotal role in the portrayal of realism, depth, and aesthetics.

- **Importance of Accurate Lighting Capture:**
  - *Realism:* Lighting provides visual cues about the shape, texture, and depth of objects. Capturing this accurately is paramount to ensuring realistic virtual recreations.
  - *Mood Setting:* Lighting influences the tone and emotion of a scene. For example, dim lighting can create an intimate or suspenseful mood, while bright lighting can imply joy or clarity.
- **Offline Capture Techniques:**

  The following summarizes some techniques to allow for the capturing of scene lighting as part of a pre-production phase. These can then be used as the static lights of a scene.

  - *Light Probes:* Spherical devices that capture surrounding light from all directions. These are especially useful for capturing the interplay of light in an environment and are crucial for virtual reality or CGI where real-world lighting conditions need to be replicated in a digital space.
  - *High Dynamic Range (HDR) Imaging:* This captures a greater range of luminosity than standard digital imaging. An HDR camera can capture both the darkest and brightest parts of a scene, allowing for a full spectrum of light to be utilized in post-production. Since the screens of VR devices are typically Standard Dynamic Range (SDR), techniques like *tone mapping*[9] must be used to map the HDR values to SDR.
  - *Reflectance Fields:* Capturing how light interacts with different surfaces, whether it's the sheen of an actor's skin or the texture of a fabric, is essential. This information allows digital artists to recreate materials with high fidelity.
- **Lighting Control Systems:**
  - *DMX Console*: A device used to control lighting and effects in real-time. It's a part of the production setup and is responsible for sending control data to lighting fixtures, fog machines, and other controllable hardware.

[9] Mantiuk, R., Daly, S., & Kerofsky, L. (n.d.). Display Adaptive Tone Mapping.

- **Challenges in Lighting Capture:**
  - *Dynamic Environments:* Lighting conditions can change rapidly, especially in outdoor settings. This demands swift adaptations and robust capture systems.
  - *Complex Surfaces:* Different materials reflect and absorb light in various ways. Capturing these nuances is challenging but vital for accurate recreation.
  - *Shadow Interplay:* Shadows are a direct outcome of lighting and can be a part of a production. Accurately capturing the softness, direction, and depth of shadows is essential for realism and aesthetics.
- **Integration with Motion Capture:**
  - *Consistency:* When capturing both motion and lighting, it's crucial to maintain consistency. A slight change in lighting can influence the perceived motion or depth especially for systems that rely on visual data.
  - *Post-Processing:* After capture, the lighting data can be utilized to light digital avatars or objects. This integration ensures that the digital representation benefits from the real-world lighting conditions during the performance capture.

In summary, the process of capturing lighting is both complex and detailed. But with the right techniques and careful attention, it can significantly elevate the authenticity and quality of performance capture, guaranteeing a result that is visually captivating and emotionally impactful.

3.3. Props

Capturing the movement and interaction of props adds another layer of complexity and detail to a performance. Whether it's a sword in a combat scene or a ball in a dance sequence, props can significantly influence the narrative and emotion of a performance. A very important prop in many productions is clothing which is a non-rigid body, and it is quite challenging to capture. Here's what to consider in the capturing of props:

- **Capture Systems:**
  - *Optical Tracking Markers:* Placing optical markers on props enables motion capture systems to track them in real-time.
  - *RFID Tags:* These can be used for less dynamic props to record when they move in and out of a scene.
- **Key Considerations for Prop Capture:**
  - *Synchronization*: Like audio, props need to be accurately synchronized with the performer's movements and with other elements like lighting and audio.
  - *Interactions*: The way a prop interacts with a performer or other props needs to be captured accurately. This could be as simple as a hand grabbing a glass or as complex as two swords clashing.

- *Soft body props, like clothing* are notoriously difficult to capture. A common approach for clothing is to use cloth simulation to dynamically simulate the dynamics of clothing as it interacts with the performers and the environment.
- *Multi-Prop Management*: Some performances may involve multiple props being used simultaneously. Each must be individually tagged and tracked to ensure accurate data capture.

- **Challenges in Prop Capture:**
  - *Occlusion*: When a prop is hidden from the capture system's cameras, it can result in a loss of tracking data.
  - *Size and Shape Variability*: Props come in all shapes and sizes, which can be a challenge for capture systems that are optimized for human forms.
  - *Material Properties*: The properties of the prop material, like reflectiveness or colour, can affect how well it is captured.
  - *Soft-body props, like clothing* are quite challenging to capture. Volumetric capture systems are able to capture everything together, including clothing appearance and dynamics; these captures are not easily editable. In all other situations, the best approach is to simply do cloth simulation. This also enables the use of different clothing by the same performer in a virtual setting.
- **Integration with Other Modalities:**
  - *Data Fusion*: Prop data can be combined with motion, audio, and lighting data to create a more complete and realistic capture.
  - *Post-Processing*: Captured prop data may require post-processing to adjust for any errors or to add additional detail.

In summary, the capturing of props, although challenging, is a necessary aspect in some performances that contributes to the overall richness and complexity of a performance. With the correct setup and integration, props can become as lifelike and interactive as the performers themselves, adding depth to the captured experience.

4. Fundamental Criteria in selecting Performance Capture systems

Selecting the right performance capture system is a critical decision that impacts the quality, efficiency, and overall success of a project. To make an informed choice, several fundamental criteria must be considered. Each of these criteria plays a unique role in determining which system aligns best with specific project requirements.

4.1. Budget

Budget considerations are foundational when selecting a set of performance capture systems. The financial resources available for the project significantly influence the options and capabilities of the chosen systems. Evaluating budget constraints helps in finding a solution that delivers optimal value without compromising project goals.

When evaluating costs, it's essential to look beyond just the initial purchase price and delve deeper into the long-term financial implications (e.g., maintenance, licenses, technical expertise needed, etc.) and the value you will receive. Here are some key aspects to consider:

- *Initial Purchase Price:* This is the upfront cost of the hardware and software. Depending on the technology and its capabilities, prices can vary significantly. Typically, the more accurate a solution, the more expensive it is.
- *Software Licenses:* Some performance capture systems require ongoing licenses for the software used to process and analyze the data. These licenses might be one-time fees, annual subscriptions, or monthly charges.
- *Maintenance and Upgrades:* Over time, all technology requires maintenance. Cameras or sensors might need replacement parts. Software might require upgrades to remain compatible with other systems or to gain new features.
- *Training Costs:* If the team is unfamiliar with the technology, there might be training costs involved. These can be in the form of official training sessions, or the time cost associated with self-learning and experimentation.
- *Operational Costs:* Some systems might consume a lot of power, require specialized facilities, or need frequent calibration. These operational costs can add up over time.
- *Expansion and Scalability:* If you anticipate needing to expand your system in the future, consider the costs associated with adding more cameras, sensors, or software licenses. Some systems might be more cost-effective to scale than others.
- *Resale Value:* Technology can become outdated quickly. If you're considering an upgrade in a few years, it might be worth checking the potential resale value of the system you're buying. Some brands and technologies retain their value better than others.

4.2. Intrusiveness

The level of intrusiveness associated with a performance capture system is a crucial consideration. Some systems may require performers to wear sensors, markers, or specialized attire, impacting their freedom of movement and comfort. Evaluating the level of intrusiveness ensures that the chosen system aligns with the performers' needs and the project's objectives.

4.3. Number of performers

The scale of a performance, including the number of performers involved, influences the choice of a capture system. Some systems are optimized for single performers or small groups, while others excel in capturing large ensembles. Determining the number of performers is vital in selecting a system that can effectively handle the intended scope of the project. Additionally, if one uses different systems that exhibit drifting (e.g., XSens and Rokoko), long capture sessions might lead to data overlap or inconsistent motions especially in close interactions between performers.

4.4. Multi-performer localization

Localizations refers to the placement of each performer in the correct place in the common performance stage; this is especially crucial in an XR environment. *On-site vs. Remote performance:* Some systems require performers to be on-site in a specific environment (like a studio), while others can capture performances remotely. The latter implies that the performers must be placed by someone in the correct places in a common virtual space.

4.5. Tracking accuracy

Performance capture can focus on capturing pose (the orientation and configuration of body parts), positions (the spatial location of objects or performers) or both. This criterion is vital for projects where precision in either pose or position data is essential. Understanding this distinction aids in selecting a system that aligns with project objectives. Outside-in systems with markers are typically the more accurate solutions, whereas inside-out systems or purely vision-based systems without any markers are less accurate in both orientation and 3D localization.

4.6. Do you need facial tracking?

Facial expressions and emotions are integral aspects of many performances. Assessing a performance capture system's ability to capture facial movements and expressions is critical for projects where facial fidelity is essential, such as character-driven animations for movies or theatre or clinical assessments. This typically requires different software and hardware (e.g., head mounted rig, stereo and/or depth cameras) to body capturing. There are companies like Rokoko that offer a complete solution for body, face and hands.

Figure 4.1 Facial tracking is typically vision based and might be helped with markers to capture facial landmarks.

4.7. Do you need hands and fingers tracking?

The level of detail and precision required for hand and finger movements varies across projects. Certain applications demand high-fidelity hand and finger capture, while others may prioritize broader body movements. Evaluating a system's capability to capture hands and fingers helps tailor it to project-specific needs. Several mocap systems provide specialized gloves to capture finger motion; most typical systems only track body poses.

4.8. Interactive (performance, collaboration) vs Off-line (analysis, storage)

Performance capture systems may serve different purposes, ranging from real-time interactive performances to offline analysis and data storage. Distinguishing between interactive and off-line requirements is vital to select a system optimized for the intended usage, whether it involves live performances, collaborative experiences, or post-production analysis and archiving. In live performances for example, the capture systems should be fast, responsive, accurate and provide a high-speed network interface with low latency to stream the data.

4.9. Portability

Portability can be a vital aspect of motion capture systems, enabling their versatile use across industries and environments such as sports science, biomechanics, healthcare, entertainment, and robotics. In the performing arts, it facilitates on-location tracking and can contribute to immersive experiences by adapting to users' environments such as a theatre or stage.

5. Requirements in the context of PREMIERE

The following section summarizes requirements a) for the four use cases as defined in the PREMIERE project and b) per type of performance (dance, theatre). *A more detailed and thorough description of the use cases can be found in D2.1 (User Requirements).*

5.1. Requirements per Use Case

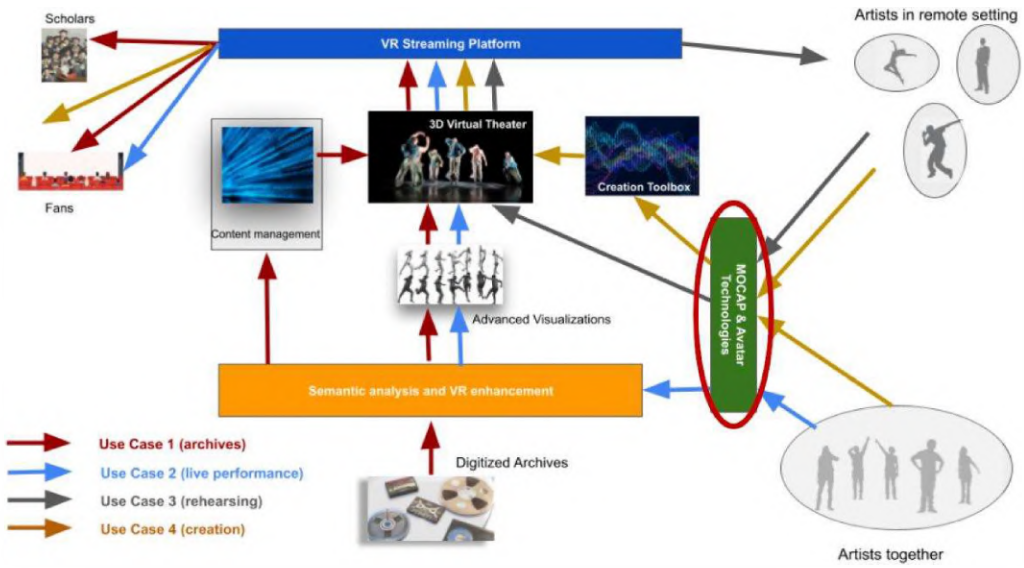

Figure 5.1 Semantic relation of use cases, technologies, and reusable components. The motion capture (mocap) system is highlighted with a red circle.

In the context of PREMIERE, there are four distinct use cases depicted in Figure 5.1. Performance and motion capture is employed in use cases 2, 3, and 4, which relate to live performances, rehearsals, and creative processes, respectively. However, data capture is also necessary for use case 1. Early in the project, the use case leaders were asked to provide the requirements for their respective use cases; these included functional and technical requirements such as number of performers, the type of performance, what needs to be tracked and recorded, etc. In Table 5.1, we summarize the requirements for each of the use cases.

Requirements description	Use case
Capture dance movement with home camera equipment.	1
Record rehearsals and performances across multiple modalities (video, audio, motion capture, wearable sensors).	1,3
Track positions and trajectories of performers on stage in relation to one another and the scene/space.	2,3,4
Track changing area that the performer's body occupies.	3
Track poses and movements of performers' full bodies and hands	2,3,4
Track interaction and synchronization between performers	3,4
Track position and movement of stage props, static and moving objects	2,3

Track hand gestures of performers	2,3,4
Track facial expressions of performers	2,3,4
Detect scene changes and events in scene	1,2,3
Detect properties and changes in music	1,2,3
Detect lighting conditions and light movements.	1,2,3
Focus on capturing systems that are unobtrusive and comfortable if wearable	3
Provide gesture-based controls for capturing	3

Table 5.1 Functional requirements for the four use cases.

5.1.1. Use Case 1: Performing Arts Archives browsing enhanced with AR/VR technologies

This use case focuses on analysing existing video archives using AI and 3D technologies. Performance capture is used in this use case to document and archive new performances. Depending on the effectiveness of the technology, videos may be a complete 360 VR recording or be enhanced as "legacy" content with XR features (i.e., played back in the virtual theatre on a virtual display). Such content will then be projected in the 3D virtual theatre through the VR media platform. For performance capture reasons, it is essential to record dance movements using camera equipment, be able to identify scene changes and events, detect properties and changes in music, and monitor lighting conditions and movements.

5.1.2. Use Case 2: Live performance enhanced with VR technologies

Similar to the first use case, performers will be presented within a VR replica of the physical venue through the 3D Virtual Theater. But in this case, spectators will view this in real-time via the VR streaming platform, with added real-time semantic data visualization capabilities. For this use case it is needed to capture at least performers' movements and gestures as well as any moving props. For the case of a live theatre performance, face, and speech is very important and should be captured. Additionally, loose clothing can be of importance to many performances. All these elements must be captured and streamed live on the VR platform, thus high-quality real-time tracking is crucial.

5.1.3. Use Case 3: The Virtual co-creation and rehearsing space

In the third use case, there is an emphasis on virtual performances employing a 3D theater, motion capture, and avatar technologies. However, the distinction lies in its focus on real-time interaction to co-created virtual performances. There are a few interesting technical challenges; network lag and synchronization of the actions of participants are crucial and will affect the real-time interaction of participants if not handled properly, and facial capture is difficult considering that users will be wearing VR headsets. This use case requires real-time high-quality performance capture with transmission of minimal amounts of data to address networking issues between participants.

5.1.4. Use Case 4: Artistic creation using AI and VR technologies

The fourth use case, while similar to the second and third, focuses on users involved in the creation and performance of new dance using AI and VR tools that are based on dance movements. These users can either collaborate remotely from various locations or be co-located in the same venue. We utilize motion capture systems to track the performers' movements, hand gestures, and facial expressions.

5.2. Dance vs Theater

Performance capture for theatre and dance shares similarities in the sense that they both involve capturing live human performances, but there are also significant differences between the two which depend on movements, technology to be used, artistic intent, environment, audience engagement and post-production.

Artistic Intent and Purpose

- **Theater:** In theater, performance capture is often used to document, archive, or stream live theatrical productions, enabling live distribution, recording of a play or musical for future reference, analysis, or promotion. Additionally, it can be used for co-creation and remote rehearsal purposes.
- **Dance:** In dance, performance capture can serve as a means to create choreographic records, preserve dance works, and analyze movement for artistic development and educational purposes. It is more focused on capturing the nuances of dance techniques and expressions.

Movement and Choreography

- **Theater:** The movement in theater productions can vary widely, from naturalistic acting to stylized movements and dance sequences. Performance capture in theater may prioritize capturing dialogue and character interactions.
- **Dance:** In dance, the focus is primarily on capturing the precise choreography and movements executed by dancers, with an emphasis on body lines, timing, and expression. Many times, these movements can be very fluid and dynamic, therefore body postures are fast responsive times are important.

Technology and Equipment

- **Theater:** Theater performance capture often involves the use of traditional video cameras and audio recording equipment, as well as multiple camera angles to capture the overall production. Motion capture equipment can be used to record precise movements and facial expressions.
- **Dance:** Dance performance capture may employ specialized motion capture systems with high-speed cameras, markerless tracking, or even inertial sensors to capture the intricacies of dance movements. However, whilst talking with dance students in a workshop for a PREMIERE meeting, we realized that some dancers could actually utilize the fallacies and limitations of motion capture systems to augment their performances, e.g., they would think of creative ways of exploiting glitches such as loss of tracking or drifting.

Space and Environment

- **Theater:** Theatrical productions take place on stage with various sets, props, and lighting, and performance capture aims to reproduce the live theater experience.
- **Dance:** Dance performance capture can occur in a studio or stage environment, and the focus is on capturing the pure essence of dance movements, often with minimal stage elements.

Audience Engagement

- **Theater:** Performance capture for theater is typically geared toward a wide audience, including those who may not have seen the live production. It can serve as a promotional tool and a way to reach a broader audience.
- **Dance:** Dance performance capture is often used for educational and artistic purposes within the dance community, focusing on dance technique, training, and preservation.

Post-Production and Editing

- **Theater:** Theater performance capture may involve editing and post-production work to create a polished recording for distribution or archival purposes. In the case of PREMIERE, this can involve processing by multiple systems to get semantic metadata and new data like actors, subtitles, movement analytics, etc.
- **Dance:** Dance performance capture also requires post-production editing , typically to enhance the quality of captured data (given that glitches themselves are not part of the performance as discussed a few paragraphs back). In the case of PREMIERE, this can also involve running some algorithms to extract qualitative characteristics of motion or classification of motion, etc.

5.2.1. Requirements for Dance Capture

Dancer Dynamics:

- Track the changing area occupied by the dancer's body.
- Capture full-body movements, hand gestures, and (maybe) facial expressions.
- Dancer movements are very often very fluid and dynamic.
- Monitor poses and detailed movements of dancers, including intricate hand movements.

Dancer Interactions: Track interaction and synchronization between multiple dancers, ensuring harmony and coordination in group performances.

Props & Environment:

- Track the position and movement of props used during the dance.
- Detect changes in lighting conditions and track light movements, essential for setting the mood during a dance performance.
- Simulate loose clothing dynamics instead of capturing.

Sound & Music: Detect changes and properties in music, crucial for understanding the rhythm and pacing of the dance.

Recording Equipment: Capture systems for covering multiple modalities like video, motion capture, and wearable sensors.

Wearable Equipment: Ensure any wearable sensors are non-intrusive, allowing dancers free movement.

Control Mechanisms: Implement gesture-based controls for capturing systems, facilitating seamless operation during performances.

5.2.2. Requirements for Theatre Capture

Stage Dynamics:

- Track positions and trajectories of performers on stage in relation to one another and the surrounding environment.
- Detect scene changes and events within the scene.

Actor Interactions:

- Monitor interaction and synchronization between performers.
- Capture hand gestures and facial expressions of performers.
- Track positions, poses, and movements of performers' full bodies and hands.

Props & Environment:

- Track the position and movement of stage props, both static and moving objects.
- Detect lighting conditions, shifts, and movements of lights.
- Simulate loose clothing dynamics instead of capturing.

Recording Equipment: Capture systems for covering multiple modalities like video, audio, motion capture, and wearable sensors.

Control Mechanisms: Provide gesture-based controls for capturing systems, allowing for intuitive and efficient operation.

Sound & Music: Detect properties and changes in music to understand the ambiance and mood changes.

Wearable Equipment: If wearable sensors are used, they should be unobtrusive and comfortable for performers.

6. Protocol for Performance Capture

As shown in Figure 1.1, performance capture in PREMIERE involves a) data collection of performers and their environment from the different physical locations, b) local synchronization of these data, c) transmission of data in standardized formats through a network to the virtual theatre state synchronization server which then d) ensures all performers experience the same state and e) storage of appropriate data into the CMS. In the following paragraphs we analyze how data should be captured from each local performance space.

6.1. Capturing Protocol for a performance

In the following section we describe a basic performance capture protocol that can be used for most of the use cases of the project, for both dance and theatre. This is the first version that acts as a basis for capture; *it will be revised, revisited, and adapted over the course of the project.*

The objective of this protocol is to establish a standardized and systematic approach to capturing live human performances with a focus on theater and dance to achieve the desired requirements. In Figure 6.1, the diagram illustrates the protocol for capturing a performance. Below, we explain each of the steps to follow.

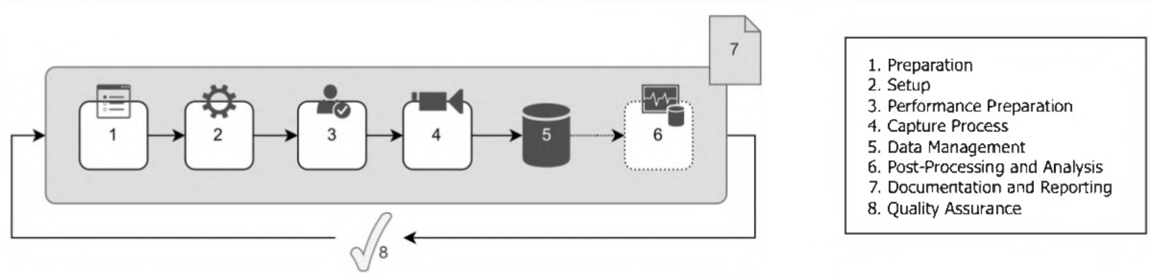

Figure 6.1 Capturing Protocol diagram for a performance

1. Preparation:
   - Define the purpose of the performance capture, whether it's for archival, analysis, real-time streaming or rehearsal.
   - Identify the type of performance (theater or dance) and its specific requirements.
   - Select the appropriate capture technology and equipment, such as motion capture system, cameras, microphones, and sensors.
2. Setup:
   - Determine the performance space and its dimensions, ensuring it meets the technical requirements of the capture system.
   - Position cameras, microphones, or other sensors according to predetermined angles and locations to capture the entire performance area.
   - Calibrate and test all equipment to ensure they are functioning correctly.
3. Performance Preparation:
   - Coordinate with performers, directors, and technical staff to ensure everyone is aware of the capture process.
   - Perform a rehearsal or blocking session to familiarize performers with the capture setup and any technical constraints.
   - If possible, ensure costumes, props, and lighting are adjusted for optimal capture.
4. Capture Process:

- Begin recording or capturing the performance according to the established guidelines.
- Monitor the capture process in real-time to identify and address any technical issues or anomalies.
- Ensure all performers and technical crew adhere to the predefined performance cues and timing.

5. Data Management:
   - Establish a data management system to organize and store the captured performance data securely.
   - Assign unique identifiers to each recording session for easy retrieval and tracking.
   - Back up all data to prevent data loss.
   - Add Universal Resource Locators for the data and add them in the Content Management System
6. Post-Processing and Analysis (if applicable):
   - If the protocol involves analysis, specify the steps for data processing and analysis, including software and tools to be used.
   - Detail how the data will be annotated, synchronized, and prepared for further study.
7. Documentation and Reporting:
   - Maintain comprehensive documentation of the capture process, including equipment used, camera angles, performer information, and any issues encountered. Use video and audio recordings if possible as video and audio blogs of the process.
   - Prepare a final report summarizing the results, findings, and recommendations for future capture sessions.
8. Quality Assurance:
   - Implement quality checks at various stages of the protocol to ensure data accuracy and integrity.
   - Establish a feedback loop for continuous improvement of the protocol based on lessons learned from previous captures.

6.2. Input: Capture Devices

A typical performance capture setup will include motion capture systems, audio recording equipment, and lighting control, all connected to several workstations possibly connected through a Local Area Network (LAN) (see Figure 6.2). Devices should communicate through a LAN (through a set of interconnected routers) to minimize latency and maximize throughput. A typical setup would include one workstation for Motion Capture and one for rendering on a VR device locally. The motion capture devices would send tracking data to the capture workstation through the LAN, which in turn would stream it in appropriate format to the rendering workstation. A real-time engine like Unreal Engine might also be part of the setup for rendering and streaming. *During the implementation phase, a setup (or a few different setups) will be defined in the context of PREMIERE and will be described in version 2 of the performance capture protocol deliverable (D2.8).*

Our recommendation for most performance capture sessions are as follows:

- *Motion Capture*: Xsens or Rokoko for body and hand movements.
- *Face:* iPhone or high-resolution cameras and/or specialized facial capture systems such as Faceware.
- *Audio:* High-quality microphones for capturing sound. MIDI output for devices that support it.

- *Lighting:* DMX-controlled lights that can be triggered or adjusted in real-time.
- *Props:* Various sensors or markers attached to props for tracking. Preferably the same motion capture system (e.g., XSens sensors attached to props) or ArUco/QR code for pose estimation.
- *Networking:* one or two routers to connect all devices in a LAN. High speed internet connection to the internet (at least 100Mbps, ideally fiber with minimal latency) – this will be tested thoroughly.

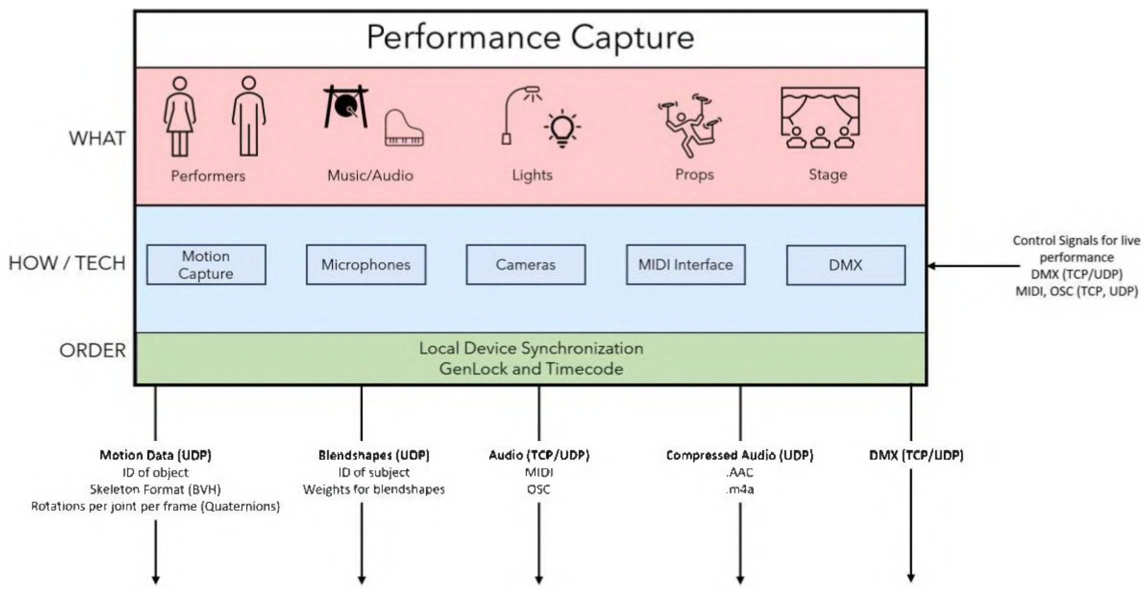

Figure 6.2 Performance capture overview diagram demonstrating the three layers of decisions that need to be made; what is to be tracked, how and how these multimodal data are to be synced together before being sent to remote users (both performers and viewers).

6.3. Output: for storage and streaming

The different devices used during performance capture are able to generate data in different formats. Raw data will be streamed to the Virtual Theatre using the appropriate LiveLink plugins of Unreal. However, for storage we should ensure that we use data formats that are as bare-bone as possible which allow for easy transcoding between other formats, especially for Machine Learning tasks. One of the fundamental problems in Machine Learning research is the collection of large amounts of properly formatted data, and having a structured way early on for collecting and describing these data in the context of PREMIERE can have a large impact. For PREMIERE especially, data are important for use cases 1 and 4; analysis of archived content and training data for data / theatre creation respectively.

6.3.1. Skeletal Data: Body and hands

Skeletal animation, also known as rigging or skeletal rigging, is a widely used technique in computer graphics and 3D animation that enables the realistic movement and deformation of 3D characters or objects. It involves the use of a hierarchical skeleton structure (also known as a rig) and associated control mechanisms to animate the model.

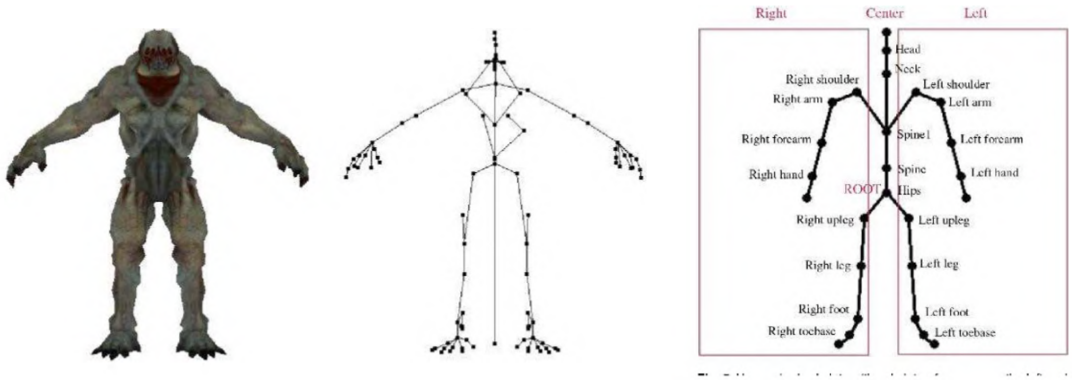

Figure 6.3 A character and its skeletal rig[10]. Image on the right shows the common BVH hierarchy

With respect to motion capture, skeletal animation requires the following components:

- Skeleton Structure (Rig): A skeletal structure is created or assigned to the 3D model, typically in the form of a hierarchical tree-like structure composed of bones or joints. Each bone or joint is connected to others, forming a hierarchy. This hierarchy represents the structure of the character or object.
- Vertices and Weights (Skinning): The mesh of the 3D model is associated with the skeleton by assigning a set of vertices (points in 3D space) to specific bones. Each vertex can be influenced by one or more bones, and the degree of influence is determined by weight values. This process is known as skinning.
- Joint Rotation and Translation: Animators can manipulate the rotation and translation (movement) of individual joints or bones within the skeleton to pose the character or object. These transformations affect the associated vertices, deforming the mesh accordingly. These are typically stored as Quaternions or rotation matrices.
- Inverse Kinematics (IK): Inverse kinematics is a technique that allows animators to control the end-effector (e.g., a character's hand or foot) of a limb, and the rest of the limb's joints automatically adjust to reach a specified position. This simplifies the animation process for natural movements.

The *BVH (Biovision Hierarchy) file format*[11] is a widely used format for storing and representing motion capture data, particularly in the context of skeletal motion. It was developed by Biovision, a motion capture hardware and software company (now owned by Autodesk) and has become a common standard for sharing and using motion capture data in animation and game development. We believe it is the simplest format that we can use to store motion capture data, however it is not the only one.

BVH files are human-readable, making them easy to create, edit, and understand. They are ASCII-based and can be opened and manipulated with text editors or script-based tools. BVH files are commonly used in animation software and game engines to drive character animations. The format supports hierarchical structures, which means that complex skeletal systems, such as humanoid characters with multiple bones and joints, can be accurately represented. BVH files are versatile and can be adapted for various applications, including character animation, biomechanical research, and motion analysis.

Description of the BVH file format

[10] https://research.ncl.ac.uk/game/mastersdegree/graphicsforgames/

[11] https://en.wikipedia.org/wiki/Biovision_Hierarchy

- File Structure: A BVH file is a plain text file with a hierarchical structure that represents the motion of a character or skeletal model. The file consists of two main sections: the HIERARCHY section and the MOTION section.
- HIERARCHY Section: The HIERARCHY section defines the structure of the skeletal hierarchy, including bones or joints, their relationships, and their respective motion channels. The hierarchy is described using a tree-like structure, with each bone or joint represented as a node in the hierarchy. The section begins with the keyword "HIERARCHY" and includes information about the root node and its children, specifying the joint's name, rotation, and any associated channels.
- MOTION Section: The MOTION section contains the actual motion data, specifying the rotations and translations of each joint at different frames in the animation. It starts with the keyword "MOTION," followed by details about the animation, including the number of frames, frame rate (frames per second), and the BVH version. The motion data is organized into rows, with each row representing the joint angles or translations for each frame. The data may include rotations in Euler angles (e.g., Xrotation, Yrotation, Zrotation) and translations (e.g., Xposition, Yposition, Zposition) for each joint. The motion data can be specified in either degrees or radians, depending on the BVH file.

While BVH files are well-suited for skeletal motion capture data, they may not capture other types of data, such as facial expressions or finger movements. For capturing a broader range of motion data, other file formats like FBX or C3D may be more appropriate. However, for many animation and game development tasks involving character motion, BVH remains a popular and accessible choice. Recently, researchers in Computer Vision introduced the SMPL and SMPL-X formats[12] as a more complete representation of motion that includes the entire body. This is still a research platform and we propose that we do not use this as the standard for the project, however, we should support converting between BVH and SMPL as much as possible, especially considering that several Machine Learning models expect data in that format.

To conclude, we suggest that the mocap data should be stored in BVH and/or FBX for now, since they are currently the most commonly used standards in the industry and research. We should also implement/use converters for important formats if needed for specific applications.

6.3.2. Blendshapes / morph targets / shape keys: For the Face

Blendshapes, also known as morph targets or shape keys are a technique used in computer graphics and animation to create smooth and realistic deformations of 3D models, particularly character faces. Blendshapes allow animators to interpolate between a set of predefined shapes or poses to achieve a wide range of facial expressions and shapes.

The key components of blendshapes are:

- Key Poses : In a blendshape-based system, a 3D model is created with a set of key poses or shapes that represent various facial expressions, phonemes for speech, or other deformation targets such as smiling, sad, angry, etc. (see Figure 6.3). These key poses are typically manually sculpted or created through 3D modeling software.

[12] https://smpl.is.tue.mpg.de/

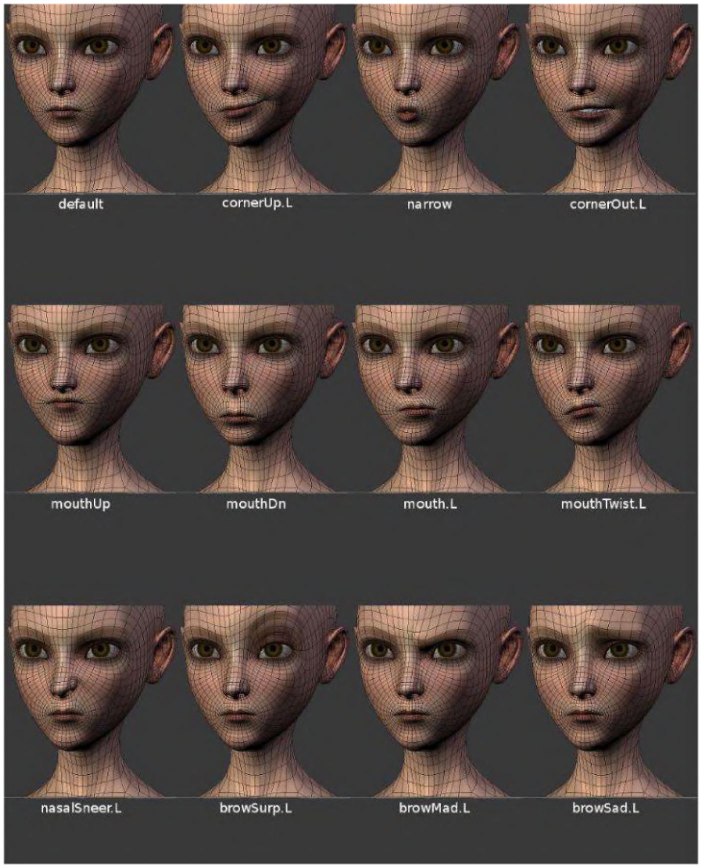

Figure 6.3 Facial blendshapes from the Blender Foundation Open-Source movie Durian[13].

- Interpolation: The magic of blendshapes lies in interpolation (see Figure 6.4). Software can smoothly transition between these key poses by blending or morphing between them. This means that, instead of manually animating every single facial expression or shape, an animator can create a range of expressions by adjusting the weights or percentages of each blendshape.

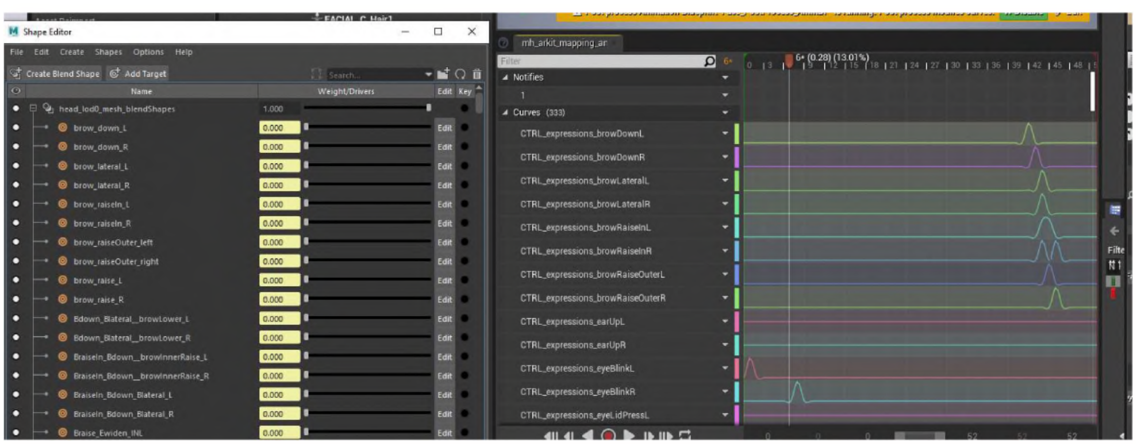

Figure 6.4 Blendshape weights between Maya and Unreal[14]. Each raw represents a key pose and the values in the timeline indicate the weight at any given time.

13 https://durian.blender.org/news/sintel-facial-blendshapes/

14 https://forums.unrealengine.com/t/metahuman-maya-blendshape-to-unreal-different-names/233026

- <u>Realistic Animation</u>: Blendshapes are particularly useful in achieving realistic facial animations, such as lip syncing, emotional expressions, and facial movements. By blending together multiple shapes, a character can exhibit nuanced and lifelike expressions without the need for complex rigging or bones.

Unfortunately, there is not a widely recognized standardized file format for blendshapes. Instead, blendshapes are typically implemented and stored within 3D modeling and animation software in a way that is specific to each software package. Each software tool may have its own method for defining, naming, and controlling blendshapes.

However, there are common file formats that are used for 3D models that may include blendshape data when the model is exported or saved. These formats can include:

- <u>FBX (Filmbox)</u>: The FBX file format, developed by Autodesk, is a widely used format for exchanging 3D models and animations between different software applications. It can store blendshape data along with the 3D model, making it a common choice for character animation pipelines.
- <u>Alembic (ABC)</u>: Alembic is an open-source 3D interchange format that is often used for efficient caching and transfer of animated geometry, including blendshape data, between different software packages.
- <u>Collada (DAE)</u>: Collada is an open standard XML-based format for 3D content exchange. It can be used to export 3D models with blendshape information.
- <u>Proprietary Formats</u>: Some 3D software tools have their own proprietary file formats that include blendshape information. For example, Maya's .ma and .mb file formats can store blendshape data created using Autodesk Maya.

In PREMIERE we will investigate which approach is better for our case and select accordingly. For now, it seems the FBX file format is the optimal choice to store the complete motion of a performer, however it is not an open-source solution and FBX are not human readable as BVH.

6.3.3. Audio (waveforms)

There are many file formats to store audio. The choice of format depends on factors like audio quality, compatibility with editing software, and intended use. Here are some standard formats commonly used to store live performance audio:

Here's a summary of the standard formats commonly used to store live performance audio:

- <u>WAV and AIFF</u>: These formats offer high-quality, uncompressed audio suitable for archiving and professional editing.
- <u>FLAC</u>: A lossless compression format that reduces file size without sacrificing audio quality, making it ideal for archiving.
- <u>MP3</u>: A widely used compressed format for online distribution, known for smaller file sizes but with some loss of audio data at lower bitrates.
- <u>AAC</u>: Another compressed format with good audio quality at lower bitrates, commonly used for online distribution, especially on Apple devices.
- <u>OGG Vorbis</u>: An open-source, efficient, and quality-compressed format suitable for streaming and gaming.
- <u>M4A</u>: A container format that can store audio in various codecs, including AAC and Apple Lossless, commonly used with iTunes and Apple devices.
- <u>WebM</u>: An open and royalty-free format primarily used for web-based audio and video.

- PCM: Uncompressed and retains original audio quality, often used for archival purposes but results in large file sizes.

The choice of format depends on specific needs. For high-quality archiving, WAV, AIFF or FLAC may be preferred. For online real-time distribution, MP3, AAC or OGG formats are commonly used, with the choice depending on the target audience and platform compatibility. It is also important to consider the bitrate and sample rate settings when encoding compressed audio formats to balance quality and file size. We will investigate which formats are appropriate for PREMIERE over the next few months.

6.3.4. Audio (OSC and/or MIDI)

One of or both of MIDI and OSC[15] should be used in PREMIERE to store data from electronic musical instruments, like synthesizers, guitars and pianos. MIDI stands for Musical Instrument Digital Interface, whereas OSC stands for Open Sound Control. Both are versatile technologies and protocols used in music production and performance. They serve as standardized languages for communication between electronic musical instruments, computers, and other digital devices.

MIDI is a communication protocol that transmits musical information, such as notes, dynamics, and control data, between devices. It is event-based, conveying discrete musical events like note-on, note-off, and control change messages. MIDI allows real-time control of musical instruments and devices, enabling musicians to play, record, and manipulate music. *MIDI does not transmit audio itself; it conveys instructions and data for music performance and control.* It is compatible with a wide range of digital devices, including synthesizers, keyboards, computers, and software applications. Finally, it is easily editable and used for creating, arranging, composing, and producing music.

OSC is a communication protocol used primarily for exchanging real-time musical and multimedia data over networks. It's designed to facilitate the transmission of control messages, sensor data, and other time-sensitive information between software, hardware, and multimedia devices. OSC is particularly popular in audiovisual performance, interactive art, music production, and multimedia installations. Key features of OSC include its flexibility, simplicity, and suitability for both local and networked communication. Instead of relying on fixed formats like MIDI, OSC allows users to define their custom data structures and message formats, making it highly adaptable for a wide range of creative and interactive applications.

[15] https://opensoundcontrol.stanford.edu/

7. VR streaming and Synchronization

PREMIERE aims to distribute content that might be generated in real-time by performers in different geographic locations to different types of users. Synchronization is a fundamental element in networked performances, ensuring that participants can collaborate effectively, deliver their artistic expressions with precision, and provide audiences with a cohesive and engaging experience. Whether it's a live-streamed dance performance, a virtual theater production, or a collaborative online art project, synchronization plays a vital role in the success and impact of networked performances.

To achieve these goals, data in PREMIERE should be synchronized at different levels; a) all devices on a physical location should be synchronized through a local network using timecode and potentially genlock, b) the virtual theatre should be perceived synchronously by performers at a global network level (i.e., the internet) and finally different users should perceive/receive a synchronized view of the performance and any relevant data.

7.1. Local Device Synchronization

To be able to connect and synchronize all the devices for the performance capture, we achieve that via Timecode and Genlock (see Figure 7.1). The use of Genlock is important because it provides not just frame-accurate synchronization via timecode, but also that all devices function with the same frequency.

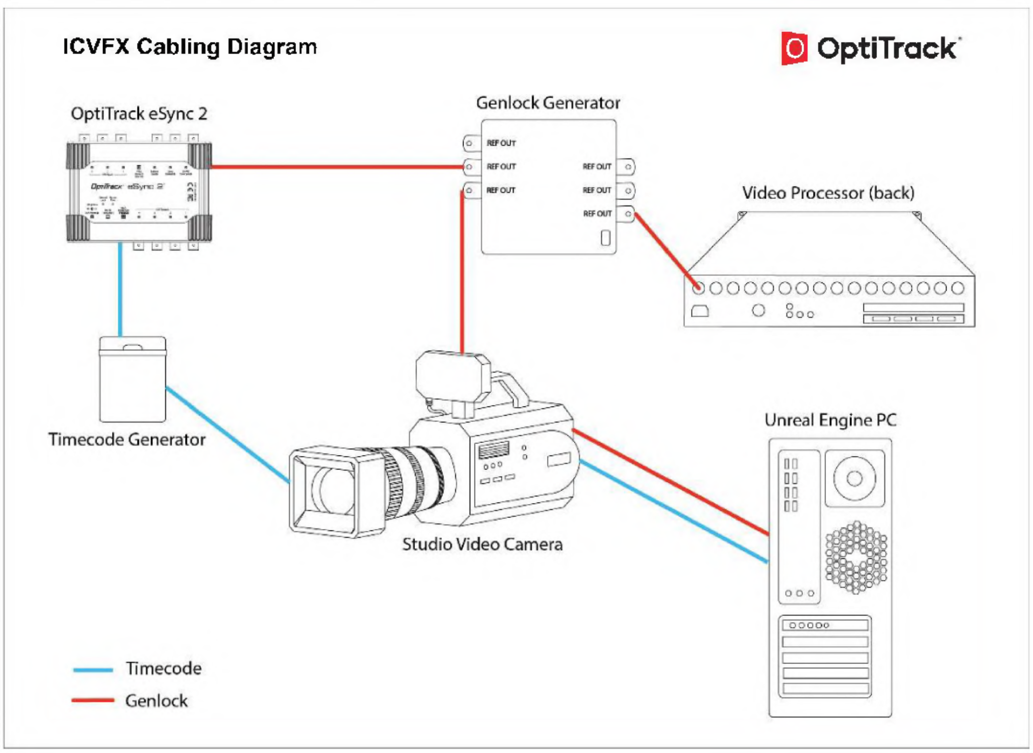

Figure 7.1 Genlock and timecode for Virtual Production using the OptiTrack Motion Capture system[16]. The GenLock signal (red) synchronizes all devices in real-time and the timecode (blue) can be used for alignment.

Most Motion Capture systems have genlock and timecode support. For example, for one of the proposed inside-out motion capture systems (Xsens), there are two options depending

16 Image from: https://docs.optitrack.com/v/v2.3/virtual-production/unreal-engine-optitrack-incamera-vfx

on if we are using the Link or the Awinda systems[17] (see Figure 7.2). We use a central control PC or a Sync unit to start and stop recordings on the different devices in the same network. In the following figures we demonstrate how this works for the two systems (images from the official XSens documentation)

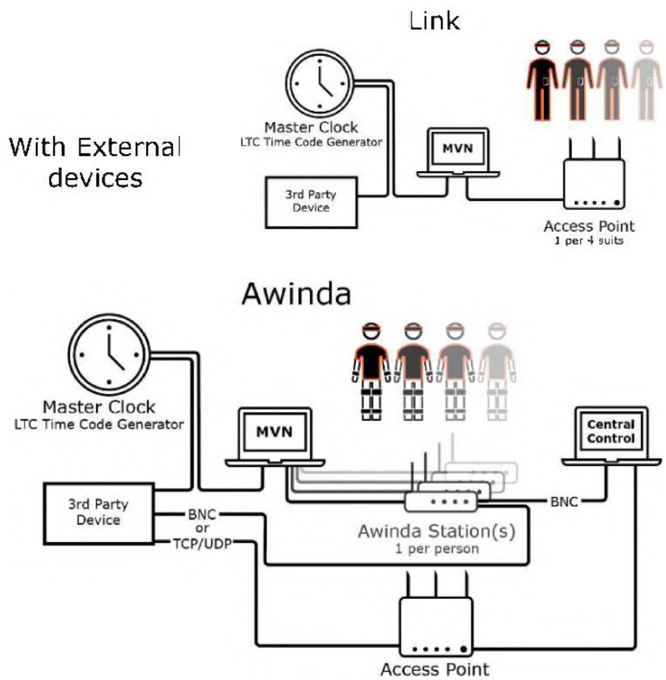

Figure 7.2 Synchronization of multiple XSens suits using timecode.

7.2. Network Synchronization

Synchronized data is important in networked performances. This will a) ensure participants act in unison despite physical separation, b) enable immediate responses to cues and prompts, enhancing audience engagement, c) preserve the artistic vision and timing critical to the performance, d) ensure seamless interactions and fewer disruptions with minimal latency, e) foster collaboration among artists from various disciplines and f) offer a consistent experience to audiences worldwide.

There are many techniques used in Networked games that can handle network synchronization and can be adapted to performances[18]. These include:

- *Server-Based Synchronization*: Use a dedicated server that acts as the authoritative source of game/performance state. All performer actions are processed and validated by the server, and the server broadcasts updates to all connected clients. All connected clients use the server clock as the basis for synchronization.
- *Latency Compensation*: Implement mechanisms to compensate for network latency. Techniques like client-side prediction and server reconciliation help minimize the impact of lag on player experiences.
- *Frame Rate Synchronization*: Ensure that all players run the game at the same frame rate to maintain consistent timing and animations. Frame rate discrepancies can

[17] https://base.movella.com/s/article/LTC-Timecode-in-MVN?language=en_US
[18] See for example: https://pvigier.github.io/2019/09/08/beginner-guide-game-networking.html

PREMiERE

lead to uneven experiences. Using CloudXR to stream content to different systems can help in this since the rendering systems will be the same across users.

- *Synchronized Clocks*: or timestamps are used to coordinate events and actions across all players. This helps in maintaining temporal consistency.
- *Shared Randomness*: When randomness is involved (e.g., random sounds, light parameters, etc.), synchronize the random seed or use deterministic algorithms to generate random outcomes. This ensures that all users have the same random experiences.
- *Network Interpolation*: Smooth out networked data such as motion capture data by applying interpolation and extrapolation techniques. This reduces the perception of jittery or choppy movement.
- *State Replication*: Continuously replicate state information about the virtual theater (e.g., performers' positions and poses) to all connected clients.
- *Replay System*: This is used in games to record game sessions; we should follow a similar approach to record a performance with all variables, random seeds, etc. This allows viewers to replay an experience.
- *Time-Synchronized Events*: Trigger events at specific times in the theatre world based on synchronized server time. This ensures that events occur simultaneously for all participants.
- *Consistent Physics*: Ensure that physics, such as collision detection and object interactions produce identical results on all clients.
- *Session Resynchronization*: Allow for resynchronization in case of network disruptions or desynchronization issues. This prevents players from being permanently out of sync.
- *Server Reconciliation*: Use techniques like dead reckoning and server reconciliation to correct discrepancies in player positions and actions that may occur due to network latency.

7.2.1. Network Synchronization in Unreal Engine

The PREMIERE Virtual Theatre is implemented using the Unreal Game Engine. Unreal supports several methods for network synchronization between clients[19]. This process involves leveraging Unreal's replication system to duplicate and synchronize crucial gameplay data across the network. Actors, which represent in-game objects and characters, are meticulously replicated to ensure that all players share a coherent view of the game world. Remote Function Calls (RFCs) are utilized to trigger synchronized events and actions across all connected clients, with special attention to RFC ordering for precise synchronization.

Player input replication, lag compensation, and server authority are essential aspects of maintaining synchronization, ensuring that player actions align correctly across networked clients. Thorough testing and debugging are crucial to identify and address synchronization issues, while optimization techniques help minimize network traffic and enhance performance. Unreal Engine's networking tools, including the Replication Graph system in recent versions, empower developers to fine-tune and optimize network synchronization, providing players with a seamless and enjoyable multiplayer gaming experience.

Current Networking Implementation in the PREMIERE Virtual Theatre: Currently, we are utilizing the preferred way for doing multiplayer experiences in Unreal which is ***Epic Online***

---

19 https://docs.unrealengine.com/5.3/en-US/networking-and-multiplayer-in-unreal-engine/

Services (EOS); this is a set of online multiplayer and networking tools and services provided by Epic Games. EOS is designed to simplify the development of online multiplayer games and enable cross-platform play, allowing players from different gaming platforms to connect and play together.

7.3. CloudXR for VR Streaming

VR streaming is a significant leap that bridges the gap between live performances and remote audiences, offering a viewing experience that surpasses the confines of a traditional screen.

Among the tools available for VR streaming, CloudXR emerges as a prominent solution. It facilitates the streaming of performances captured in real-time, making them accessible to remote audiences.

Features of CloudXR:

- *Real-time Performance Relay*: XR Cloud ensures that there's minimal lag between the actual performance and its virtual representation, providing an almost instantaneous relay of events.
- *High Fidelity Streaming*: The platform retains the quality and intricacy of the captured performance, ensuring that viewers get a crisp, high-definition experience.
- *Scalability*: XR Cloud can cater to a wide range of audience sizes, from a few individuals to large-scale global viewership.
- *Interactivity*: Viewers can choose their vantage point, interact with the environment, or even communicate with other viewers, making the experience interactive and social.

8. Conclusion

As the world of performing arts undergoes a significant transformation, embracing both its storied traditions and the exciting possibilities of the future, cutting-edge technologies like motion capture and VR streaming emerge as pivotal drivers of this evolution. Throughout the course of this document, we've delved into the intricate facets of performance capture systems, spanning from the complexities of Inside-Out and Outside-In methodologies to the multifaceted considerations that influence their selection.

In the specific context of the PREMIERE project, the unique requirements of dance and theatre take center stage. These art forms, renowned for their nuance and expressive depth, demand tools that can faithfully capture not only movement but also audio, props, lighting, and the surrounding environment. Moreover, these elements must be seamlessly transmitted to an audience, which could be located anywhere in the world. This is precisely where VR streaming assumes an indispensable role. It offers an immersive and unparalleled experience, transporting viewers directly into the heart of the performance, effectively erasing geographical distances, and ensuring that the power of art knows no boundaries.

While we've established an initial version of the protocol in this phase, it's essential to acknowledge that this version is not set in stone and won't be the final iteration used for the project. The nature of this project, which involves both technical and artistic elements, dictates that refinements and adjustments are not only expected but also necessary.

The decisions made during this initial phase were guided by Deliverable D2.1, which documented the User Requirements in detail. However, it's vital to recognize that *artistic creation is a dynamic and ever-evolving process*. What may seem like the best approach at this stage could require reconsideration in the months ahead as the creative direction of the project matures and evolves.

Furthermore, *technical challenges are an inherent part of any project*, and this one is by no means an exception; it combines several cutting edge technologies in very dynamic settings. Factors such as the intricacies of lighting and networking in different venues can significantly impact what can and should be implemented in the protocol. These technical intricacies may prompt us to revisit and adapt certain decisions made earlier.

Rest assured that all of these developments, refinements, and adjustments will be meticulously documented and presented in detail in our forthcoming deliverable, D2.6. This document will serve as a comprehensive record of the protocol's evolution, providing valuable insights into the decision-making process and how we've addressed both technical and artistic challenges to ensure the success of the project.