.

3D MOTION CAPTURED ANIMATION USING POSE DETECTION ALGORITHMS

A Research Presented to
The Faculty of the College of Engineering and Technology
University of St. La Salle
Bacolod City

In Partial Fulfillment
of the Requirements for the Subject
Thesis 1

Balbin, Karr Christopher
Canaria, Kean Gabriel
Estiamba, Kent Matthew
Sarmiento, Jarl Keenen

May 25, 2023

APPROVAL SHEET

The Thesis entitled **“3D MOTION CAPTURED ANIMATION USING POSE DETECTION ALGORITHMS”** presented by **KARR CHRISTOPHER BALBIN, KEAN GABRIEL CANARIA, KENT MATTHEW ESTIAMBA,** and **JARL KEENEN SARMIENTO,** in partial fulfillment of the requirements for the degree of Bachelor of Science in **Computer Science** of the University of St. La Salle, have been evaluated and approved by the panel of evaluators.

PANEL OF EVALUATORS

JIM JONATHAN C. DECRIPITO, PhD
Chair

NIEL G. BUNDA, PhD Member	**PAUL JOHN M. MONTAÑO** Member

EISCHIED G. ARCENAL, PhD
Adviser

ACKNOWLEDGEMENT

The researchers express their gratitude to everyone who made this research feasible. The completion of this study would not have been possible without the assistance of their thesis advisor, Dr. Eischied G. Arsenal, PhD, whose knowledge, encouragement, feedback, guidance, and unwavering support were indispensable in determining the direction of the research and ensuring that the highest standards of output were met.

In addition, the researchers express their thanks to the panel of evaluators, Dr. Jim Jonathan C. Decripito, PhD, Dr. Niel G. Bunda, PhD, and Mr. Paul John M. Montaño, for their insightful feedback and contributions throughout the research process. Their constructive criticism and recommendations assisted in refining the research and enhancing the researchers' understanding of the topic. The researchers would also like to thank Mr. Sherwin Malaga, for advising and consultation of the study as the statistician.

The faculty and students of the Information Technology and Computer Science Department are also acknowledged for their support, assistance, and participation during the beta testing.

Finally, we would like to extend our gratitude to the researchers' family and friends for their unwavering support and encouragement throughout this academic voyage. This achievement would not have been possible without their love, understanding, and patience. We trust that this work will contribute to the advancement of knowledge in the researchers' field.

DEDICATION

This thesis paper is humbly dedicated to the pursuit of knowledge and the advancement of the chosen field by the researchers. This work exemplifies the researchers' dedication and efforts to stretch the boundaries of understanding and advancing the social and technological limits of motion captured animation and pose detection algorithms.

This research is dedicated to the researchers' adviser, Dr. Eischied G. Arsenal, PhD, and the panel of evaluators, who without their support and supervision, this paper would not exist. Their confidence during the development phase until its conclusion affirms the researchers' abilities and commitment to fostering intellectual contribution and importance.

This research is also dedicated to filmmakers, animators, 3D artists, game developers, and anyone interested in 3D animation, as this research is in their field of work and expertise.

Finally, the researchers dedicate this research to the families and participants, whose support, understanding, and encouragement have sustained the researchers' confidence throughout this difficult journey. This dedication honors the collaborative spirit that drives the pursuit of knowledge and motivates future researchers to pursue excellence in their research endeavors.

ABSTRACT

This study focused on generating a 3D Motion Captured Animation Using Pose Detection Algorithms with a converter tool website that accepted a video file as input to be sent to the server for processing. A descriptive and developmental method of research was utilized in this study. The participants of the study were twenty (20) students, fifteen (15) filmmakers and animators, fifteen (15) video game developers, and five (5) IT experts, with a total of fifty-five (55) participants. An interview was first held with five (5) interviewees to determine the problems experienced by users in terms of converting video files to 3D animation files. A researcher-made questionnaire and the standardized PSSUQ questionnaire was used to measure the acceptability and usability, respectively. For the statistical tools, the mean and standard deviation was used to measure the acceptability and usability of the system. To measure the level of efficiency of the VIBE model for pose detection, an evaluation script that utilized PyTorch methods to load datasets and models was used to evaluate the models on the datasets to compare their results. The Procrustes-aligned mean per joint position error (PA-MPJPE) metric was solved, which is the 3D distance between the ground truth and the predicted joints in millimeters. The findings of the study revealed that the developed motion capture system was determined to be very acceptable by the users. It was also very usable in terms of System Usefulness, Information Quality, Interface, and Overall usability. Additionally, the VIBE model was determined to be more efficient than previous Human Pose and Shape Estimation models such as SPIN and Temporal HMR with a lesser PA-MPJPE. Finally, as the basis of the development of the motion capture system, the researchers successfully implemented the VIBE model to a converter tool website that allows users to easily convert video files to 3D animation files.

Keywords: Motion Capture, 3D Animation, VIBE, Converter, Pose Detection, Human Pose and Shape Estimation, Joint Prediction

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Motion capture has been a vital technology in various industries like video games and filmmaking. It has been used as early as the 1980s in arcade games, with films following soon after in the 1990s. It has been used to create realistic computer-generated (CG) characters and creatures such as Gollum and Smaug from adaptations of Peter Jackson's Tolkien books, The cursed crew from the Pirates of the Caribbean, the Na'vi species of Avatar, and many more. It is also used in video games such as The Last of Us, Detroit: Become Human, the Uncharted series, and God of War, to name a few.

Recently, research and development in machine learning (ML) and artificial intelligence (AI) have been rapid and many different software and solutions have emerged. These technologies have many different applications, one of which is computer vision and machine learning to detect and estimate poses made by people. Using this pose detection technology, you can record the movement of people without the use of mocap suits and complicated software. With this, smaller studios and even solo artists can create their own animations with just any camera that can record a video like their smartphone.

The researchers believed in the value of time, and that motion capture was a process that can make artists work faster, which can save them a lot of time and produce more quality works. For clarification, though, motion capture was not meant to replace traditional three-dimensional (3D) animation processes, but instead, to be an additional tool for artists to utilize when they have the need for it in their workflow. When used together with other techniques and processes, artists have more freedom in creating what they need and want.

Background of the Study

Traditional motion capture brought the advantage of producing complex and realistic movements made by actual human actors in a shorter period of time compared to traditional animation techniques, but at the cost of requiring specific and specialized hardware and software to capture and process the data. It has been utilized for everything from sports treatment to healthcare, cinema, and gaming. Studios must offer motion capture suits with markers at each joint so that the motion may be detected based on the placement and orientation of the markers. Due to the requirements of a motion capture setup, smaller or independent studios were not capable of providing said setup for clients and were left to work with only traditional animation techniques to produce an output.

With the problems stated above, introducing pose detection algorithms to the process allowed a more accessible and feasible option, even for small studios and solo artists. One such algorithm is the Video Inference for Human Body Pose and Shape Estimation (VIBE) method. Using this technology, they didn’t need to provide any specialized suits or equipment and just needed to record a standard video.

With these propositions, the researchers aimed to develop a tool that used VIBE to capture people's movements without the usage of motion capture suits or expensive software and then converted the captured data into a 3D animation file which can then be used freely by artists.

Statement of the Problem

This study aimed to develop a tool that converted video recordings to a 3D animation file. Specifically, this study sought to answer the following questions:

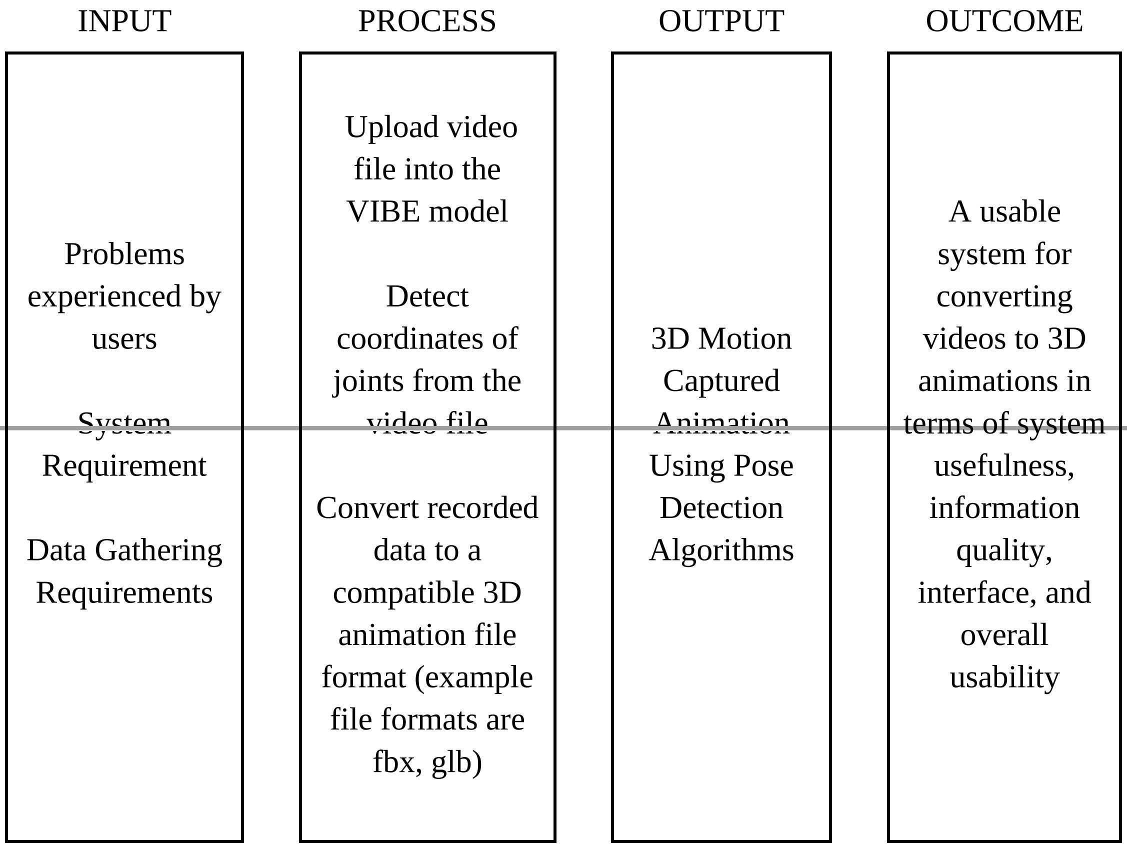
1. What were the problems experienced by the users in terms of converting video files to 3D animation files?
2. What was the level of the acceptability of the motion capture system in terms of improving the conversion of video files to 3D animation files as perceived by the users?
3. What was the usability of the motion capture system in terms of System Usefulness, Information Quality, Interface, and Overall Usability?
4. What was the level of efficiency of the pose detection algorithm?

Conceptual Framework

The conceptual framework was constructed on the IPOO (Input-Process-Output-Outcome) structure and had four stages: the input, process, output, and outcome. The IPOO diagram included every material and piece of information required for the procedure, as well as the specifics of the process itself and explanations of all products and by-products based on the process that occurred. In system analysis and software engineering, a technique was utilized to depict the configuration of a computer program or other operation, known as structural modeling (Mező, 2011).

Figure 1

Schematic Diagram Illustrating the Framework of the Study



INPUT

Problems experienced by users

System Requirement

Data Gathering Requirements

PROCESS

Upload video file into the VIBE model

Detect coordinates of joints from the video file

Convert recorded data to a compatible 3D animation file format (example file formats are fbx, glb)

OUTPUT

3D Motion Captured Animation Using Pose Detection Algorithms

OUTCOME

A usable system for converting videos to 3D animations in terms of system usefulness, information quality, interface, and overall usability

Traditional motion capture processes necessitated a large setup with motion capture suits to track the person's movements, whereas traditional 3D animation took a long time to animate each limb manually. The Skinned Multi-Person Linear (SMPL) body model requires 24 landmark points. This was where the algorithm came into play as the algorithm automatically mapped these landmarks and captured them without the need of trackers. Once the researchers received the coordinates of the landmarks from the algorithm, the data was used to create a 3D animation file.

Figure 1 showed the conceptual framework of the study, the Input phase showed the problems experienced by users, system requirements, and data gathering requirements. The Process phase showed the uploading of the video file into the pose detection machine learning model and the converter which then proceeded to its Output phase that showed the converter tool and the 3D animation file. The Outcome phase showed a finished user-friendly converter tool with its produced results from conversion.

Scope and Limitations

The study was focused on the development of a tool that captured motion from a video file and converted it into a 3D animation file. The tool’s server can be accessed locally for demo purposes through a local website using a computer or mobile browser. The server was able to process multiple inputs at the same time. When a video file finished converting, the user only needed to download the produced 3D animation file. To use the tool, the user did not need to set up an account. The tool was free for all users to use and try out. The limitation that came with not having a dedicated account was that users can’t save conversions on the cloud. This means that users can’t access an animation file that they previously converted using the tool. They had to convert the original video file again. The tool is able to detect multiple people in a single input video file and will export a zip file containing a 3D animation file for each person detected. The tool only provided a basic (a 3D-rendered human body without any clothes or accessories) 3D model for all animations when exporting the animation file because the focus of the tool was on the animation data.

Significance of the Study

The findings were significant to the following:

Animators. With the addition of this tool, animators had more options to choose from when creating animations for various projects. This tool was easily added to the workflow of an animator and the produced animation could be modified using traditional techniques using their preferred software.

Filmmakers. Independent filmmakers or studios who can’t afford a traditional motion capture setup can use this tool as an alternative that is light on the budget. It was also more accessible since it requires less physical equipment to use.

Game Developers. As with filmmakers, game developers and studios also used this tool as an alternative when creating in-game cutscenes or 3D animations for their games. This can be especially useful for small indie game developers or studios.

The Researchers. The researchers themselves benefited from this study by enhancing their expertise and refining the system for potential commercial deployment.

Future Researchers. This study also held value for future researchers as it provided them with a framework for data collection and served as a foundational component for a more comprehensive study.

.

Definition of Terms

To provide a clearer understanding of the terms used in the study, the following terms were conceptually and operationally defined.

Motion Capture. Motion capture, also known as "mocap," referred to a technology-based approach that recorded an actor's physical movements and actions in order to translate them into a computer-generated imagery (CGI) character. This advanced technique had the capability to monitor a wide range of motions, including body movements and facial expressions. While mocap was commonly used in animated films, its primary application lay in the creation of CGI characters in live-action movies. (DeGuzman, 2021)

In this study, motion capture referred to the process of capturing motion from a video file using pose detection algorithms and converting the captured data to a 3D animation file.

Machine Learning. The field of machine learning was a subset of artificial intelligence that allowed systems to enhance their performance through experience without requiring explicit programming. Essentially, machine learning involves creating computer programs that can retrieve data and apply it to learn autonomously. (Selig, 2022)

In this study, the area of research in which the creation of pose detection algorithms falls was referred to as machine learning.

Pose Detection. Pose detection was a computer vision method that involved tracking the movements of an individual or an object by detecting the location of specific key points. By analyzing these key points, the difference postures and movements were compared to gain valuable insights. Pose detection found widespread applications in various fields, such as augmented reality, animation, gaming, and robotics. (Gupta, 2021)

In this study, pose detection referred to the process that the algorithm used was responsible for doing.

Landmarks. Biologically significant points that can be accurately and consistently located were referred to as anatomical landmarks. These landmarks possessed a high degree of precision and can be identified unambiguously (Landmarks - Richtsmeier Laboratory. Available from: https://getahead.la.psu.edu/landmarks/)

In this study, landmarks were the body points that the pose detection algorithm used to help with identifying the motion of a person.

Procrustes-aligned Mean Per Joint Position Error. The Procrustes-aligned Mean Per Joint Position Error (PA-MPJPE) metric is the mean of all the euclidean distance between ground truth and prediction for a joint after the estimated 3D pose is aligned to the ground truth by the Procrustes method. (cbsudux, 2019)

In this study, the PA-MPJPE metric measured in millimeters if used to determine the level of efficiency of the VIBE model.

Ground Truth. Ground truth is a term commonly used in statistics and machine learning. It refers to the correct or "true" answer to a specific problem or question. It is a "gold standard" that can be used to compare and evaluate model results. (Datagen)

In this study, ground truth is the joint position data from the 3D datasets that were gathered through different methods. It is used to calculate the PA-MPJPE metric.

VIBE. Video Inference for Body Pose and Shape Estimation (VIBE) is a video pose and shape estimation method. It predicted the parameters of the SMPL body model for each frame of an input video. (Kocabas et al., 2020)

In this study, VIBE is the specific pose detection algorithm used in conjunction with the system.

SMPL. The Skinned Multi-Person Linear (SMPL) model is a realistic 3D model of the human body that is based on skinning and blend shapes and is learned from thousands of 3D body scans. (Loper et al., 2015)

In this study, the SMPL model is used as the base asset for the 3D animation file output generated by the system.

SPIN. SMPL oPtimization IN the loop (SPIN) is an approach that proposes a close collaboration between a regression method and an optimization-based method to train a deep network for 3D human pose and shape estimation. (Kolotouros et al., 2019)

In this study, SPIN is evaluated together with the VIBE and Temporal HMR models to compare their efficiency using the PA-MPJPE metric.

Temporal HMR. Human Mesh Recovery (HMR) is an end-to-end framework for reconstructing a full 3D mesh of a human body from a single RGB image. (Kanazawa et al., 2018)

In this study, Temporal HMR is evaluated together with the VIBE and SPIN models to compare their efficiency using the PA-MPJPE metric.

REVIEW OF RELATED LITERATURE

This section includes related concepts and related literature that the researchers consider relevant to the study. The theories, concepts, and milestones regarding the development and application of motion capture technology and pose detection algorithms is likewise discussed.

Motion Capture Technology

In the film and television industries, motion capture technology is more mature and widely used. We can obtain 3D virtual animation by capturing the motion data of professional actors, performing specific processing, and then binding it with the character model in film and television works (R. Zeng, 2021). The use of major motion capture characters has become more widespread due to advancements in technology. Nowadays, performers engaged in motion capture are mandated to utilize a complete outfit that encompasses their entire body and is adorned with the track of dots that are aligned with precise locations on a CGI model. These dots are recorded by cameras, which capture the position of the dots as well as the changing distances between them. This forms the basis for creating animation. While these suits enable filmmakers to record an actor's entire body, there are also other methods for motion tracking specific body parts. For instance, facial motion capture is used to create realistic facial animations, such as Benedict Cumberbatch's portrayal of Smaug in The Hobbit trilogy (Okoyomon, 2019).

As human-computer interaction technology advances, natural interaction and pattern recognition will become the primary means of interaction between humans and computers. To improve motion capture technology, it is essential for the computer to understand and capture the unique characteristics of human behavior. Given that the human body comprises over 200 rotary joints, to simulate human body movement realistically, it is necessary to provide the value and position of each joint's rotation angle, along with other relevant information. This paper presents a literature review on developing a motion capture system to analyze athletes' actions, and the results show the effectiveness of the proposed system. To further evaluate the system's robustness, we plan to conduct additional experiments in the future (Fang Han, Xuesong Bo, 2015).

Development of Pose Detection Algorithms using Machine Learning

In recent years, researchers both domestic and abroad have explored various convolutional neural network models and auxiliary techniques to assess human posture, ranging from single to multi-person, 2D to 3D, and photo to video (Yating Wei, 2022). The 3D positions of human joints in a global coordinate system can be detected using 2D joint locations from multi-view video camera images, similar to marker-based motion capture systems. Recent deep-learning-based computer vision research has focused on 3D pose estimation, using a single algorithm to directly determine 3D joint locations.

Studies have explored 3D pose estimation using single-view camera images by Chen and Ramanan (2017), Pavlakos (2018), and Moon (2019), as well as Rhodin(2018), Iskakov (2019), and Pavllo (2019), centered their research on the examination of human motion and body positions. However, biomechanics researchers require accuracy as well as ease of use to meet the goals of motion analysis. Seethapathi et al. (2019) evaluated pose-tracking research from a movement science standpoint and found that deep-learning-based human pose-tracking algorithms did not prioritize the quantities that matter for movement science. The accuracy of deep-learning-based 3D markerless motion capture for human movement investigations, such as sports biomechanics or clinical biomechanics, is unknown.

There are two general types of 3D pose estimation: regressive, which regresses the 3D coordinates of nodes directly from 2D data, and lifting (C. Cao et al., 2021), which obtains the two-dimensional pose first and then uses a mapping method to lift it to the three-dimensional space on top of the two.

Several datasets with 3D pose annotations in single-person settings (Mehta et al. 2017; Trumble et al. 2017; von Marcard et al. 2016) or multi-person scenarios with only 2D pose annotations (Mehta et al. 2016) exist. As multi-person 3D pose estimation gained popularity, datasets such as MarCOnI (Elhayek et al. 2016) with fewer scenes and subjects, as well as more diverse datasets such as Panoptic (Hanbyul Joo and Sheikh 2015) and MuCo-3DHP (Mehta et al. 2018) emerged. LCRNet (Rogez et al. 2017) creates faux annotations on the MPII 2D pose dataset using 2D to 3D lifting, while LCRNet++ (Rogez et al. 2019) employs synthetic representations of individuals from a variety of single-person datasets.

3D Multi-Person Pose

In the past, researchers working on recording the three-dimensional body positions of several individuals using a single camera typically used a generative approach, which involved using a trained model to calculate the 3D body and camera pose based on 2D landmarks. This approach was limited by the capabilities of existing deep learning technologies. Rogez et al. (2017) improved upon this approach by using a

Faster-RCNN technique to identify representative poses from clusters of discrete poses and then refining them. Although this method is not real-time, it can still achieve interactive frame rates on consumer hardware. Dabral (2019) used the same point of view to project two-dimensional key points onto a 3D space for each person in a scene, using a Faster-RCNN-based method with a reduced number of anchor poses. Moon et al. (2019) demonstrated the augmentation information, certainty level of key points and three-dimensional pose representation, during the "lifting" stage can significantly improve the accuracy of pose predictions. Their method involves a preliminary person recognition step, where the posture estimation network is fed with scaled picture crops of each detected subject.

According to Cao et al. (2017), a technique that produces highly accurate pose estimates can come at the cost of longer inference times, making it more suitable for offline applications than real-time ones. In addition, as the number of subjects in the image increases, so does the per-frame inference time, making it impractical for scenarios where speed is essential. Detection-based techniques, as mentioned before, generate multiple pose predictions for each person and then combine them. However, this process can be time-consuming and may result in errors such as merging poses of different individuals or missing some proposals for the same person. Furthermore, as highlighted by Moon et al. (2019), combining position estimations and repeated identifications of the identical individual can lead to potential errors and increased inference times, further adding to the complexity of the technique. Unlike some bottom-up techniques, our approach does not result in duplicate detections for the same person. In the bottom-up method proposed by Mehta et al. (2018), all persons' 2D and 3D poses in the scene are predicted using a fixed number of feature maps that can encode any number of persons present.

However, this can lead to issues when there are overlapping subjects in the scene, requiring a complicated encoding and read-out system to be introduced. The 3D encoding considers each limb and the torso as separate objects, encoding their 3D posture in the feature maps at the pixel coordinates of their corresponding 2D joints. Even in cases where individuals in a scene are partially obscured from view, this encoding method has the ability to effectively accommodate for such occlusions and ensure that a complete and accurate representation of the scene is achieved, it can fail when body components of different subjects overlap, which is also the case for the approach proposed by Zanfir et al. (2018). In their method, the 2D and 3D poses of all subjects in the scene are encoded concurrently using a limited number of feature maps.

In contrast to (Mehta et al., 2018), the approach used by this method encodes the entire 3D posture vector at all points along the complete skeletal structure, including the areas between the joints. This approach can lead to conflicts in the 3D feature space. To address this, the method uses a function to evaluate limb grouping ideas for association, and a 3D pose decoding step to extract characteristics for each limb and combine them into a 3D pose prediction. One of the main contributions of this method is that it only considers body joints with direct visual data, such as the joint itself or its parent/child, which are used to generate a compact fully-connected network to convert the available information into a complete 3D posture prediction, even if some data is missing or imperfect. By combining the image-to-pose regression and 2D-to-3D lifting techniques, our approach overcomes the limitations of using these techniques separately. When there are no conflicts, the 3D pose encodings provide a strong signal for the 3D posture, while the global context and 2D pose information help resolve any conflicts that do arise and fill in missing body joints. This is in contrast to the approaches of (Zanfir et al., 2018) and (Mehta et al., 2018) which encode the position of all joints in the entire body or limb-wise, respectively, without considering available picture evidence for each joint.

We also employ a kinematic limitations-based fitting step which improves the temporal smoothness of our predictions. While the approach of (Zanfir et al., 2018) also uses learning and optimization, their space-time optimization over all frames is not suitable for real-time applications.

3D Data Sets

Currently, the most widely used motion capture systems are mechanical, electromagnetic, acoustic, and optical technologies. Among these, optical motion capture systems are predominantly relying on the utilization of multiple cameras to record, this method seeks to comprehensively capture every angle and detail of the subject, resulting in a comprehensive and detailed documentation of sequences of motion image and trajectories. These systems then identify and track specific markers in the image data, and use the motion information of these markers to reconstruct the motion in 3D.

The Wild dataset in 3D Poses by von Marcard et al. (2018) is a dataset of people photographed outside with a moving camera and with ground truth 3D posture. However, the number of subjects with ground truth data is limited. To increase training diversity, the authors employed the MuCo-3DHP dataset (Mehta et al. 2018), which is a multi-person training set of composited actual photos with 3D posture annotations from the single person MPI-INF3DHP dataset. In terms of convolutional network designs, ResNet variants (Xie et al. 2017) integrate explicit information from earlier to later feature layers in the network via summation skip links, which enables the development of deeper and more powerful networks. However, deeper and more powerful networks come at the expense of longer calculation times during inference and a larger number of parameters. Therefore, specific designs for quicker test time computation and parameter efficiency, such as AmoebaNet (Real et al. 2019), MobileNet (Real et al. 2019), ESPNet (Mehta et al. 2018), ERFNet (Romera et al. 2018), EfficientNet, and SqueezeNet (Iandola et al. 2016), have also been proposed.

Several convolutional architectures are tailored to perform well on specific edge devices, often at the expense of accuracy. To improve accuracy, some of these architectures may require an increase in network width or depth, which can lead to comparable GPU runtimes to typical ResNet architectures. This approach was highlighted in a study by Howard (2019) and Sandler (2018).When it comes to posture estimation, the presence of grid artifacts can negatively impact part association performance. To address this issue, ShuffleNet Zhang (2018) incorporates grouped convolutions, depth-wise separable convolutions, also shuffled as feature maps or activation maps to encourage feedforward or inference across feature groups. Densely connected convolutional networks Huang (2017), on the other hand, uses complete residual connections to create a lightweight network. However, the computational cost of concatenation operations results in sluggish performance due to the significant memory usage. Recent studies have introduced computationally efficient networks, such as those proposed by Sun et al. (2019) and Wang et al. (2019), that maintain high-to-low resolution feature representations throughout the network without sacrificing accuracy. However, in practical settings, these approaches do not necessarily result in faster computational speeds.In fact, these models may perform doubling as slowly as residual networks at predetermined accuracy threshold. Methods that prioritize model efficiency may not always lead to performance improvements due to high computational costs or inefficient sparsity-inducing transformations.These issues have been documented in prior studies by Iandola et al. (2016) and Frankle and Carbin (2018), respectively.

Dilated convolutions are often used to increase the receptive field of a network without increasing the number of parameters or sacrificing resolution. However, as mentioned in the previous statement, they can result in non-smooth output maps with grid artifacts. These artifacts can have a negative impact on part association performance in posture estimation tasks, as they make it more difficult to accurately identify and track specific joints or body parts. Therefore, it is important to carefully consider the use of dilated convolutions in such applications, and to explore alternative approaches that can achieve the desired receptive field without introducing unwanted artifacts. The ShuffleNet architecture, introduced by Zhang and colleagues in 2018, improves information flow between different groups of channels in a convolutional neural network using a combination of group convolutions, depthwise convolutions, and shuffled channels. In contrast, the DenseNet architecture, developed by Huang and colleagues in 2017, achieves parameter efficiency by using dense concatenation-skip connectivity, but this approach requires a lot of memory and can result in slower performance.

There have been recent developments in creating networks that are efficient in terms of computation while maintaining accuracy by preserving high to low resolution feature representations throughout the network. However, despite the theoretical gains in computational efficiency, these models tend to be slower than ResNet networks by up to twice the speed at the same accuracy level. Efforts to make neural network models more parameter-efficient often do not lead to significant improvements in computational speed because the computational cost may still be high or the non-structured sparse operations resulting from weight pruning may not be efficient on current hardware (Choi Jong-In, 2019).

The use of 3D estimation is especially beneficial for detecting multiple individuals. To extract essential information from the video stream, time-domain convolution is employed to determine the 3D position at different intervals. Graph convolution is necessary to implement a graph structure that treats the relationship between human key points as a graph, which helps retrieve the 3D bone relationship using a combined global and local approach. Precisely estimating 3D human position from 2D joint locations is crucial for analyzing images and videos of people. While many methods use priors to solve the ill-posed problem of human pose estimation, these priors may not consider how joint limits fluctuate with position, resulting in incorrect poses. Our work offers two significant contributions. Firstly, we create a motion capture dataset that covers a wide range of human poses, allowing us to develop a pose-dependent model of joint limitations that serves as our prior (Akhter and Black, 2017). Secondly, we propose a method that incorporates joint-limits in the priors to produce more accurate and realistic poses.

Motion Capture System

Mechanical motion capture systems rely on mechanical devices to track and record the motion trajectory. Mechanical motion capture systems are often made up of many joints and robust connecting rods (Gu, X. et al., 2016). When the gadget is in motion, the location and trajectory of the rod end point in space may be calculated using the angle change registered by the angle sensor and the length of the linkage. The mechanical tracking systems rely on machinery instruments to follow and record movement paths. These platforms are usually made up of numerous jointed elements and resilient connecting bars (Gu et al., 2016). As the mechanical device moves, an angle sensor detects alterations in angle, and the length of the linkage is measured to ascertain the endpoint of the rod's location and trajectory in three-dimensional space.

An acoustic motion capture system is composed of acoustic motion, a processing unit, a transmitter, and a receiver. The stationary ultrasonic generator serves as the transmitter, while the receiver comprises three triangular ultrasonic probes (Y. Tong et al., 2021). Although relatively inexpensive, this technology suffers from significant latency and lag, poor real-time performance, and generally low accuracy. Additionally, large obstructions between the sound source and receiver can cause interference, noise, and multiple reflections. The latter method must also take into account the correlation between air pressure and the velocity of acoustic propagation (C. Mo et al., 2021).

The motion capture system utilizing optical technology is widely employed and considered to be the most convenient system globally. This system employs various infrared cameras in recording objects from multiple viewpoints. Captured images are evaluated by a program to determine pixel locations of the tracking markers in the picture. Computer vision techniques are then utilized to conduct 3D reconstruction and obtain the kinematic data of the motion capture markers. This method of Mocap offers the broad scope of performance actions, has the limitless from cables or mechanical devices, is user-friendly, and offers a high sampling rate (D. Zhou et al., 2019).

Synthesis

The related literature provides relevant information on the process of motion capture technology and the development of pose detection algorithms using machine learning. After many years of improvements, motion capture technology has matured and is now a standard option for some projects in the film and video game industry.

Meanwhile, research and development on neural network models for many different machine learning topics such as human pose estimation, face detection, object detection, and many others have also been successful with different products available for open-source usage. Specifically, research on pose detection algorithms has reached many different milestones. Earlier studies worked with similar ideas as traditional motion capture technology where the identifying human body landmarks were done using markers. Further studies developed markerless technology using deep-learning-based computer vision algorithms to identify landmarks.

The researchers are confident in developing a tool to convert videos to 3D animation files using the discussed related literature as a basis, guide, and inspiration. The technology is heavily studied by many researchers and results from many different papers are abundant. Studies focusing on pose detection algorithms using machine learning usually produce the model as a standalone solution. It is the responsibility of developers to implement the solution in their own products.

DESIGN AND METHODOLOGY

In this section, the specifics of the research design, as well as the respondents of the study, instruments to be used, data gathering procedures, statistical tools, and the ethical considerations to be used in the development of the system were discussed.

Research Design

In this study, the researchers utilized the descriptive and developmental methods of research. Gillaco (2014) discussed how the descriptive method aims to uncover accurate information regarding a present circumstance by concentrating on depicting, contrasting, scrutinizing, and construing available data. Essentially, this approach centered on gathering objective details and providing a comprehensive account of the current situation. It is used to describe, explain, validate, and evaluate the results of the system acceptability survey given to respondents. This means that after conducting this research, an output was created.

The purpose of this research was to determine how to convert a video to a 3D animation, which focused on a specific design, development, or process while also identifying the conditions that allowed for its successful use.

Participants of the Study

The respondents of the study included students, aspiring filmmakers and animators, IT experts, video game developers, and 3D artists. At least five participants with experience in working with 3D animation or motion capture systems were interviewed to determine the problems experienced by users in converting video files to

3D animation files. Fifty participants tested the system in terms of its acceptability and those same fifty participants were also surveyed to determine the usability of the system.

Firstly, The IT experts and professionals with animation and game developing experience as respondents of the study were those who utilized computer-generated imagery that is either to be mixed with live-action footage or to be purely animated using digital techniques to produce films of any length. Five IT experts were interviewed to test the validity of the system and fifteen filmmakers or animators tested the system.

Secondly, video game developers as respondents were those who were primarily developing 3D games with humanoid characters. Fifteen video game developers tested the system.

Lastly, students that ranged from high school students, senior high school students, and college students who often uploaded or would like to try producing content related to 3D animation were also considered to be participants of the study. Twenty students tested the system.

Table 1

Frequency and distribution of the Respondents

Group	***f***	**%**
Filmmakers and Animators	15	27.27
Video Game Developers	15	27.27
Students	20	36.36
IT Experts	5	9.10
Total	55	100

Table 1 shows the frequency and distribution of the respondents. The total number of respondents was fifty-five. There were fifteen filmmakers and animators, which was 27.27% of the total respondents. This was also the same number with the video game developers. The twenty students were 36.36% of the total number of respondents. Lastly, the five IT Experts were 9.10% of the total number of respondents.

Instruments

To determine the problems experienced by users in converting video files to 3D animation files, interviews were conducted with people with experience in working with 3D animation or motion capture systems. The interviews were conducted with introductory questions about the respondent such as their occupation, if they worked with 3D animation in their occupation or as a hobby, how often they worked with 3D animation or motion capture systems, etc. The interviewer then asked about the problems experienced by the respondent regarding converting motion capture data or video files to 3D animation.

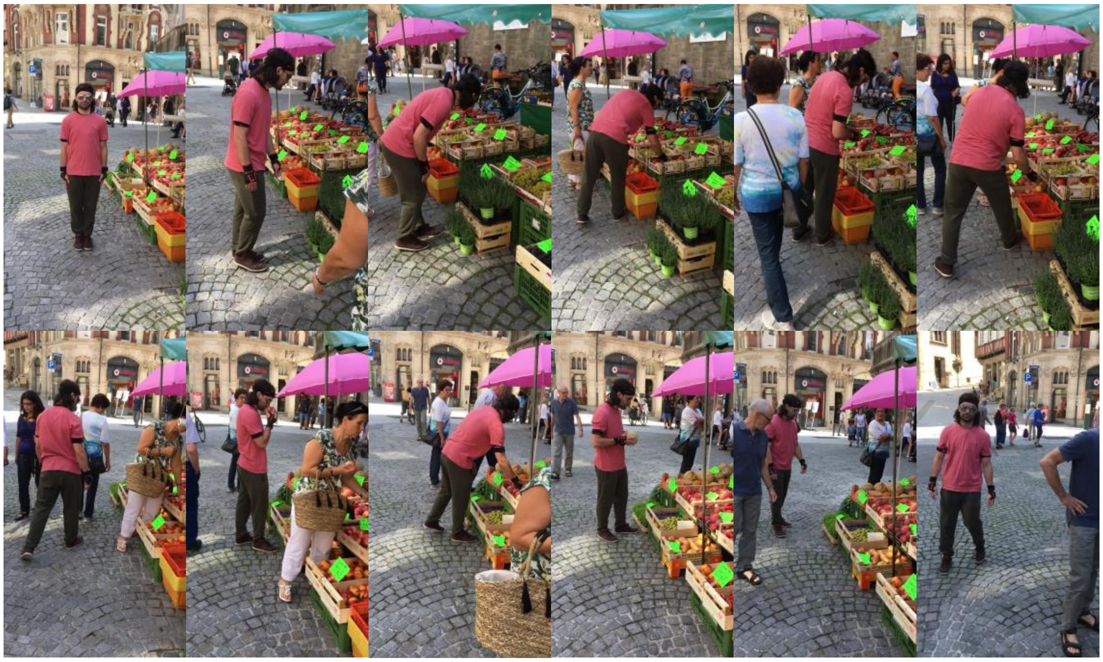
To assess the level of the acceptability of the motion capture system in terms of improving the conversion of video files to 3D animation files, an acceptability questionnaire was distributed to the participants of the study. The questionnaire had different questions aimed at determining whether users accepted the system.

To measure the usability of the system, a questionnaire that uses the PSSUQ standard as the basis of the evaluation of the software quality was provided. The usability questionnaire is divided into two components. Part I of the study instrument collected the respondents' profiles but was only optional, while Part II of the study instrument comprised a checklist sheet with 16 items.

To determine the level of efficiency of the pose detection algorithm, a python evaluation script written by the authors of the VIBE model was used. A model can be chosen through the evaluator configuration. The evaluator loads a specified dataset using PyTorch and then runs the model on the loaded dataset.

Figure 2

Sample images from the 3DPW dataset



The evaluator then calculates the PA-MPJPE metric using the model’s predicted joints and the ground truth from the dataset. The VIBE model was evaluated together with previous models, SPIN and Temporal HMR, for comparison.

Validity of the Data Gathering Instrument

To test the validity of the acceptability questionnaire, a survey instrument validation form by Carter V. Good and Douglas B. Scates was distributed to experts in their respective fields. Five experts answered the Good and Scates validation form and the mean rating was used for interpretation. The mean rating of the results of the validity survey from the five experts is 42.2, which was interpreted as Excellent by the statistician.

Reliability of the Data Gathering Instrument

To test the reliability of the acceptability questionnaire, the measure used was the Cronbach’s Alpha coefficient. This coefficient was used to measure the internal consistency of survey instruments and questionnaires.

Table 2

Interpretation Table for the Cronbach’s Alpha

Cronbach’s Alpha	Internal Consistency
α ≥ 0.9	Excellent
0.9 > α ≥ 0.8	Good
0.8 > α ≥ 0.7	Acceptable
0.7 > α ≥ 0.6	Questionable
0.6 > α ≥ 0.5	Poor
0.5 > α	Unacceptable

In the interpretation table found in Table 2 above, a Cronbach’s Alpha of 0.9 or above means the instrument has an excellent internal consistency. A Cronbach’s Alpha below 0.9 but is equal to or above 0.8 is interpreted as having a good internal consistency. A Cronbach’s Alpha below 0.8 but is equal to or above 0.7 is interpreted as having an acceptable internal consistency. A Cronbach’s Alpha below 0.7 but is equal to or above 0.6 is interpreted as having a questionable internal consistency. A Cronbach’s Alpha below 0.6 but is equal to or above 0.5 is interpreted as having a poor internal consistency. And lastly, a Cronbach’s Alpha below 0.5 is interpreted as having an unacceptable internal consistency.

Table 3

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	No. of Items
.782	.790	8

Table 3 shows the reliability statistics of the instrument which has 8 items. The results show a Cronbach’s Alpha coefficient of 0.782, which is interpreted as acceptable based on the interpretation table found in Table 2. The Cronbach’s Alpha coefficient, if based on standardized items, is 0.790, which is also interpreted as acceptable based on the interpretation table found in Table 2.

Table 4

Item Statistics

VAR00001			
VAR00002			
VAR00003			
VAR00004			
VAR00005			
VAR00006			
VAR00007			
VAR00008	4.6667	.47946	30

Table 4 shows the item statistics of the instrument. The table provides the mean and standard deviation of each item in the instrument with a sample size of 30 for all items.

Table 5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
VAR00001	32.0667	6.133	.561	.419	.752
VAR00002	32.3333	5.816	.467	.387	.761
VAR00003	32.4667	5.292	.542	.430	.750
VAR00004	32.1667	6.144	.377	.328	.775
VAR00005	32.4333	5.426	.552	.374	.746
VAR00006	32.3333	5.885	.440	.282	.766
VAR00007	32.0667	6.340	.451	.535	.765
VAR00008	32.2000	5.890	.563	.436	.747

Table 5 shows the Item-Total Statistics of the instrument, which contains, for each item, the scale mean if item was deleted, the scale variance if item was deleted, the corrected item-total correlation, the squared multiple correlation, and the Cronbach's Alpha if the item was deleted.

Data Gathering Procedures

The data from the interviews for people with experience in working with 3D animation or motion capture systems and the acceptability questionnaire was compiled and summarized in its own document for better analysis.

The acceptability questionnaire had a physical copy to be distributed to the respondents. Afterwards, the qualitative data gathered from the survey questionnaire was also compiled and summarized for better analysis.

The usability questionnaire also had physical copies that were distributed to the respondents. The questionnaires were then collected and tallied. The data was transferred to a spreadsheet for analysis.

Statistical Tools

The acceptability questionnaire used was an 8-item questionnaire with a 5-point Likert scale. The questionnaire enabled users to indicate feedback as to the overall system acceptability in terms of improving the conversion of video files to 3D animation files.

Table 6

Level of the acceptability of the motion capture system in terms of improving the conversion of video files to 3D animation files as perceived by the users

Verbal Interpretation	**Range of Mean Score**
Very Low	1.00 - 1.80
Low	1.81 - 2.60
Average	2.61 - 3.40
High	3.41 - 4.20
Very High	4.21 - 5.00

The mean score and standard deviation of each item on the acceptability questionnaire was utilized. To interpret the mean, the mean-range table found in Table 7 was used. The more items with a higher mean score (moving towards “Very High”) means that the respondents feel that the system is acceptable.

The usability questionnaire used was a 16-item questionnaire with a 7-point Likert scale and a Not Applicable (N/A) option. The questionnaire allowed users to indicate responses regarding the utilization and testing of the system.

Table 7

Hypothetical Range with Interpretation for Usability

Range of Mean Score	Verbal Interpretation
1.00 - 1.85	Very Low
1.86 - 2.71	Low
2.72 - 3.57	Moderately Low
3.58 - 4.43	Neutral
4.44 - 5.29	Moderately High
5.30 - 6.15	High
6.16 - 7.00	Very High

The mean score and standard deviation of each item on the usability questionnaire was utilized. To interpret the mean, the mean-range table found in Table 6 was used. The more items with a higher mean score (moving towards “Very High”) means that the respondents feel that the system is usable.

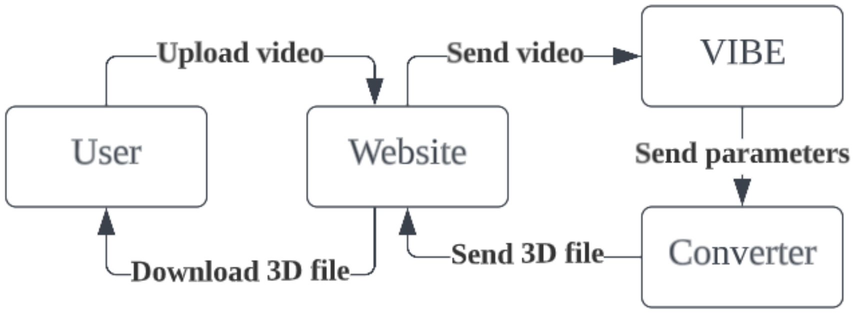
The level of efficiency of the pose detection algorithm was measured using the Procrustes-aligned mean per joint position error (PA-MPJPE) metric in millimeters. The PA-MPJPE is the mean of all the euclidean distance 2614 between ground truth and prediction for a joint after the estimated 3D pose is aligned to the ground truth by the Procrustes method. A lower PA-MPJPE meant that the predicted pose was more accurate.

System Design

The overall system functionality as indicated herein, indicates a brief description of the modular functionality of the program. The input of the system began with a user uploading a video file to the website. The website then sent the uploaded video file to the VIBE model to predict the pose of the person/people in the video.

Figure 3

Systems Operational Design of the Study



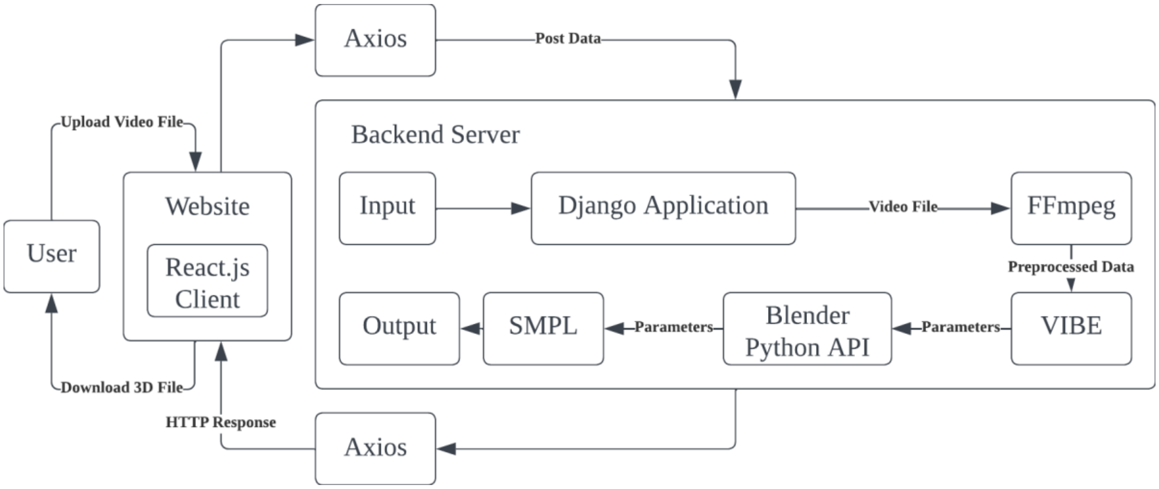
After processing the video, the parameter data was then passed to the converter which generated a 3D animation file. The 3D animation file was then sent back to the website in the HTTP response for the user to download.

System Implementation

The website was built using React.js. The user can upload a video file on the website which was then sent to the backend server via a POST request. The backend server was built with Django as a web application framework. After receiving the video file, it was then preprocessed using FFmpeg where individual frames were extracted. The frames were then passed to the VIBE model to generate the 3D data.

Figure 4

System Architecture of the Proposed Video to 3D Converter Tool



The framework of VIBE started with using a pretrained Convolutional Neural Network (CNN) which extracted the features of each frame of an input video. A temporal encoder composed of bidirectional Gated Recurrent Units (GRU) was then trained to output latent variables that contained information from past and future frames. These features were then used to regress the parameters of the SMPL body model for the whole input sequence. To enforce the pose generator to produce realistic and feasible poses, a motion discriminator was trained using a dataset called AMASS. If the pose generator was able to fool the motion discriminator, then the pose was realistic (Kocabas et al., 2020). The 3D data was then used as parameters for the SMPL 3D model asset. The asset was imported to the Blender Python API (BPY) and the parameters applied to the asset. It was then exported as a 3D file. The 3D file was then sent back in the HTTP response to the website for the user to download.

Software Specification

The specification dictated the necessary components of the system. This section enumerates all the packages, libraries, or frameworks that were used in the development of the system:

- React.js - A JavaScript library for creating single-page applications. It allowed building components and utilizing hooks for interactivity.
- Axios - A promise-based HTTP Client for node.js and the browser. It simplified sending HTTP requests and handling responses.
- Django - A high-level Python web framework for creating web applications. Django was used as a backend server to easily incorporate the VIBE model and converter Python scripts.
- FFmpeg - A command-line tool for handling video, audio, and other multimedia files and streams. It was used to extract each frame from the input video to be processed individually.

- VIBE - Video Inference for Body Pose and Shape Estimation (VIBE) is a video pose and shape estimation method. It predicted the parameters of the SMPL body model for each frame of an input video. (Kocabas et al., 2020)
- Blender Python API - A Python module that can be imported to access Blender's data, classes, and functions.
- SMPL - The Skinned Multi-Person Linear (SMPL) model is a realistic 3D model of the human body that is based on skinning and blend shapes and is learned from thousands of 3D body scans. (Loper et al., 2015)

Hardware Specification

The specification dictated the hardware used in running the backend server.

Laptop

- Processor: Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz, 2808 Mhz, 4 Core(s), 8 Logical Processor(s)
- RAM: 8.00 GB
- GPU: NVIDIA GeForce GTX 1050
- Operating System: Microsoft Windows 10 Home

Testing

Alpha Test

During the alpha testing, the researchers conducted internal testing of the system with five Test Cases. Test Cases are shown in Tables 8 to 12. This information was then used to make the necessary modifications and improvements to the system. This allowed the researchers to detect and correct faults early in the research process, saving both time and resources and boosting overall research quality.

Test Cases

Table 8

Test Case 01: Single Person

PROJECT: Single Person	**Test Case ID: 001**
TEST STARTED: April 24, 2023	
MODULE TO TEST: Single Person	TESTER: Kent Matthew Estiamba

DATE	ACTIVITY REQUIRED TO TEST	EVENT	EXPECTED RESPONSE	RATING	ACTUAL RESPONSE
April 24, 2023	Single Person Conversion	Upload video file with only a single person in the frame	The server returns a 3D animation file in a .fbx format	Satisfactory	The server returned a 3D animation file in a .fbx format

Table 8 showed the test performed for uploading a video file that only has a single person in the frame. The test started and was completed on April 24, 2023. After uploading the video file, the server successfully converted and returned a 3D animation file in a .fbx format.

Table 9

Test Case 02: Multi Person

PROJECT: MULTI PERSON	**Test Case ID: 002**
TEST STARTED: April 24, 2023	
MODULE TO TEST: Multi Person	TESTER: Kent Matthew Estiamba

DATE	ACTIVITY REQUIRED TO TEST	EVENT	EXPECTED RESPONSE	RATING	ACTUAL RESPONSE
April 24, 2023	Multi-Person Conversion	Upload video file with multiple people in the frame	The server returns a zip file that contains the 3D files in a .fbx format	Satisfactory	The server returned a zip file that contained the 3D animation files in a .fbx format

Table 9 showed the test performed for uploading a video file with multiple people in the frame. The test started and was completed on April 24, 2023. After uploading the video file, the server successfully converted and returned a zip file that contained the 3D animation files in a .fbx format.

Table 10

Test Case 03: Simultaneous Upload

PROJECT: SIMULTANEOUS UPLOAD			**Test Case ID: 003**		
TEST STARTED: April 28, 2023					
MODULE TO TEST: Simultaneous Upload			TESTER: Jarl Keenen Sarmiento		
DATE	ACTIVITY REQUIRED TO TEST	EVENT	EXPECTED RESPONSE	RATING	ACTUAL RESPONSE
April 28, 2023	Simultaneous Upload	Upload two video files at the same time	The server returns the corresponding zip file/.fbx file to each client that uploaded a video	Satisfactory	The server returned the corresponding zip file/.fbx file to each client that uploaded a video

Table 10 showed the test performed for simultaneously uploading video files to the server. The test started and was completed on April 28, 2023. After uploading two separate video files from two separate clients, the server successfully converted and returned a zip file that contained the 3D animation files in a .fbx format to the first client and a 3D animation file in a .fbx format to the second client. It was able to handle two concurrent requests at the same time.

Table 11

Test Case 04: Resolution

PROJECT: RESOLUTION	**Test Case ID: 004**
TEST STARTED: May 4, 2023	
MODULE TO TEST: Resolution	TESTER: Jarl Keenen Sarmiento

DATE	ACTIVITY REQUIRED TO TEST	EVENT	EXPECTED RESPONSE	RATING	ACTUAL RESPONSE
May 4, 2023	Uploading a 480p video	Upload video file with a resolution of 480p	The server returns a 3D animation file in a .fbx format	Satisfactory	The server returned a 3D animation file in a .fbx format
May 4, 2023	Uploading a 720p video	Upload video file with a resolution of 720p	The server returns a 3D animation file in a .fbx format	Satisfactory	The server returned a 3D animation file in a .fbx format
May 4, 2023	Uploading a 1080p video	Upload video file with a resolution of 1080p	The server returns a 3D animation file in a .fbx format	Satisfactory	The server returned a 3D animation file in a .fbx format

Table 11 showed the tests performed for uploading video files that have different resolutions. The resolutions tested were 480p, 720p, and 1080p. The test started and was completed on May 4, 2023. After uploading video files of different resolutions, the server successfully converted and returned a 3D animation file for each video uploaded.

Table 12

Test Case 05: Video Length

PROJECT: VIDEO LENGTH	**Test Case ID: 004**
TEST STARTED: May 6, 2023	
MODULE TO TEST: Video Length	TESTER: Jarl Keenen Sarmiento

DATE	ACTIVITY REQUIRED TO TEST	EVENT	EXPECTED RESPONSE	RATING	ACTUAL RESPONSE
May 6, 2023	Uploading a 10 second video	Upload a 10 second video file	The server returns a 3D animation file in a .fbx format	Satisfactory	The server returned a 3D animation file in a .fbx format
May 6, 2023	Uploading a 30 second video	Upload a 30 second video file	The server returns a 3D animation file in a .fbx format	Satisfactory	The server returned a 3D animation file in a .fbx format
May 6, 2023	Uploading a 1 minute video	Upload a 1 minute video file	The server returns a 3D animation file in a .fbx format	Satisfactory	The server returned a 3D animation file in a .fbx format

Table 12 showed the tests performed for uploading video files that have different lengths. The lengths tested were 10 seconds, 30 seconds, and 1 minute. The test started and was completed on May 6, 2023. After uploading video files of different lengths, the server successfully converted and returned a 3D animation file for each video uploaded.

Beta Test

The beta test was held at the University of St. La Salle on May 10, 2023. The booth was set up in front of the Montelibano College Library where participants were able to test the system. An android phone and a laptop that was connected to a monitor were provided for the participants. During the beta testing, the participants were provided with clear instructions on how to use the software, as well as any known issues or limitations of the software. The participants were then asked if they were willing to be video recorded using the android phone provided or using their own device. After capturing their desired video, it was copied to the researchers' laptop. The video was then uploaded to the system to be converted and downloaded with the output file imported into Blender to display the result. After testing, the participants were asked to complete the acceptability and the usability questionnaires.

Ethical Considerations

During the research and development of this study, the researchers ensured the safety and privacy of the participants during the whole process from testing the system to the evaluation. The system was also evaluated and developed to be secured and protected from potential mishandling and abuse, such as indecent uploaded content.

Taking these concerns into account, consent from the participants was obtained prior to involving them to ensure that the participants were aware of the study's goals and what they will be doing and that they are participating of their own free will. Securing participant information and data was prioritized in order to maintain their anonymity and confidentiality.

Participants had the option to withdraw from the study at any time if they had any problems, were uncomfortable, or were unable to continue for any other reason. This did not influence their contributions to the study, as well as their involvement and relationships with any of the researchers.

This took into consideration various principles that will guide the researchers throughout this study. These principles are as listed below:

Voluntary participation. The participants willingly participated of their own free will without any persuasion from the researchers. They were also free to withdraw at any time in case of problems or conflicts.

Informed consent. By giving proper information and details regarding the study, the participants should be able to make an informed decision as to whether they will participate or not.

Anonymity. To keep the privacy of the participants, their identity and personal information not important to the study were not gathered and recorded.

Confidentiality. All the data and evaluations gathered were kept safe and confidential. To stay true to the principle of anonymity, any personal information of the participants was not included in the research paper.

Physical safety. It is important that the participants were safe, and no harm fell on them at any point in time while being involved in the study.

Transparency. Data and information related to the study that concerned the participants was properly disseminated and was not kept from them unless for specific reasons such as that it was part of the process of the study or that the information was given on a future date.

No plagiarism. This was to ensure that the researchers were not copying someone else's work or study as their original work, with or without consent, by using it in the study without referencing or acknowledging the original author. This study was tested and approved using Turnitin software.

RESULTS AND DISCUSSIONS

This section contained a summary of the research study's results and discussed the conclusions, implications, and recommendations of the researchers. This section included the description of the system and its technical features as well as the analysis of the data that were surveyed during the beta test to evaluate the level of the acceptability of the motion capture system in terms of improving the conversion of video files to 3D animation files and the usability of the motion capture system in terms of System Usefulness, Information Quality, Interface, and Overall Usability. The results of the python evaluation script, which evaluated the VIBE, SPIN, and Temporal HMR models on the 3DPW, MPI-INF-3DHP, and Human3.6M datasets are also discussed. The results of said script were used to determine the level of efficiency of the pose detection algorithm (VIBE) used in the system by comparing the Procrustes-aligned mean per joint position error (PA-MPJPE) metric measured in millimeters of the three models.

Technical Features of the Motion Capture System

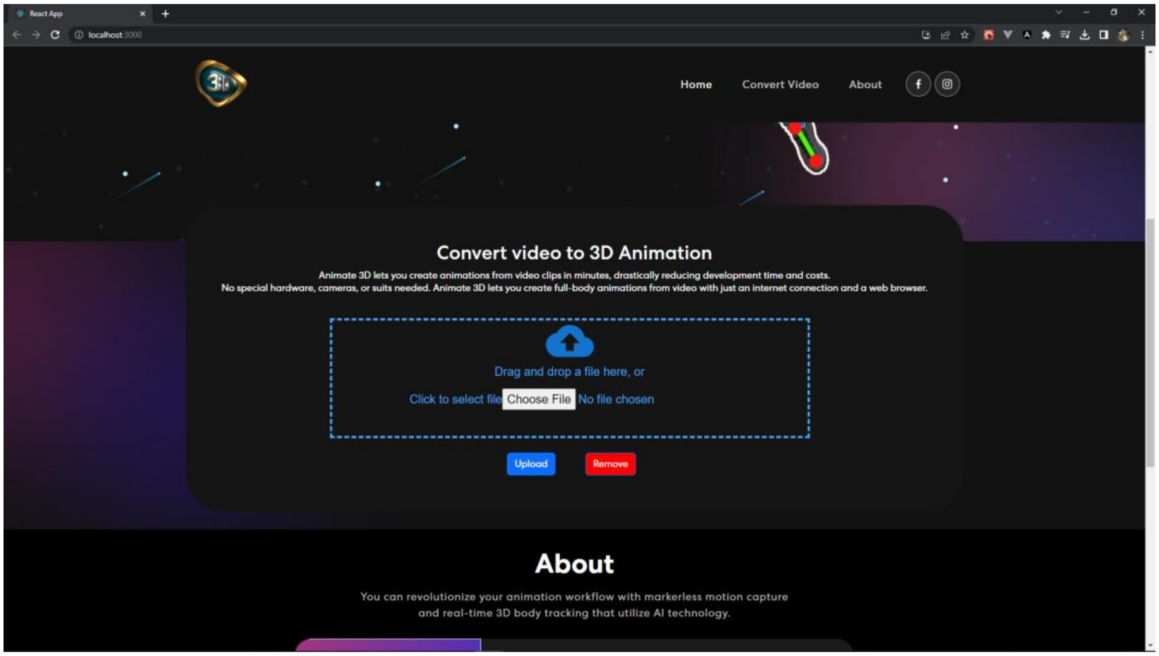
The following were the technical features of the Motion Capture System.

Upload Section of the website

This page was the entry for the users to the system and where they can upload their videos for conversion. The Upload Section provided a streamlined and informative user experience. The Upload Section was created to be noticeable for easy direction and access. Figure 5 below shows the upload section.

Figure 5

Upload section of the website


React App
localhost:3000
Home
Convert Video
About

Convert video to 3D Animation

Animate 3D lets you create animations from video clips in minutes, drastically reducing development time and costs.
No special hardware, cameras, or suits needed. Animate 3D lets you create full-body animations from video with just an internet connection and a web browser.

Drag and drop a file here, or
Click to select file Choose File No file chosen
Upload
Remove

About

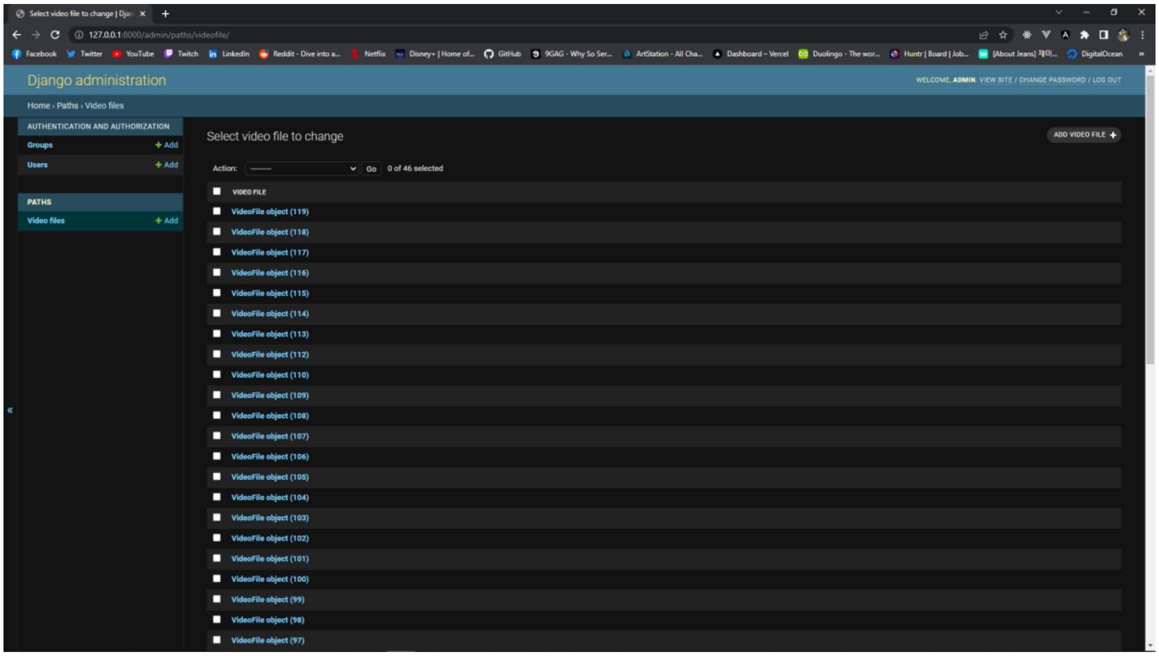
You can revolutionize your animation workflow with markerless motion capture and real-time 3D body tracking that utilize AI technology.

Django Application

The Django application exposed an API endpoint that was used by the website for uploading the input video file. It acted as the backbone of the backend server and the application that handled communication between different parts of the system such as the website, the VIBE model, the converter, and the database. Figure 6 below shows the Django administrator interface where you can interact with your database models without having to write a UI.

Figure 6

Django administrator interface



Django administration

Home › Paths › Video files

AUTHENTICATION AND AUTHORIZATION

Groups + Add

Users + Add

PATHS

Video files + Add

Select video file to change

ADD VIDEO FILE +

Action: --------- Go 0 of 46 selected

□ VIDEO FILE

□ VideoFile object (119)

□ VideoFile object (118)

□ VideoFile object (117)

□ VideoFile object (116)

□ VideoFile object (115)

□ VideoFile object (114)

□ VideoFile object (113)

□ VideoFile object (112)

□ VideoFile object (110)

□ VideoFile object (109)

□ VideoFile object (108)

□ VideoFile object (107)

□ VideoFile object (106)

□ VideoFile object (105)

□ VideoFile object (104)

□ VideoFile object (103)

□ VideoFile object (102)

□ VideoFile object (101)

□ VideoFile object (100)

□ VideoFile object (99)

□ VideoFile object (98)

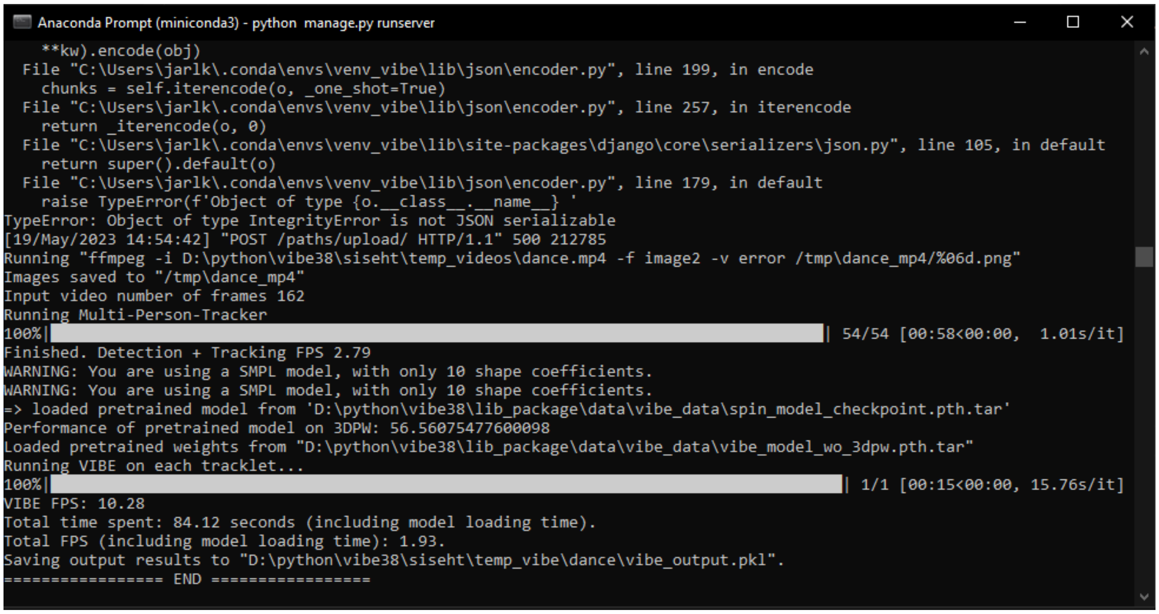
□ VideoFile object (97)

Preprocessor and Pose Detection Model

A script ran the FFmpeg tool to preprocess the input video file into individual frames which were then fed to the VIBE model which ran on every frame. The script had other various configurations such as generating a video with the 3D VIBE output embedded on top of the input video to name one. Figure 8 below shows the script running in the terminal.

Figure 7

FFmpeg and VIBE script running in the terminal



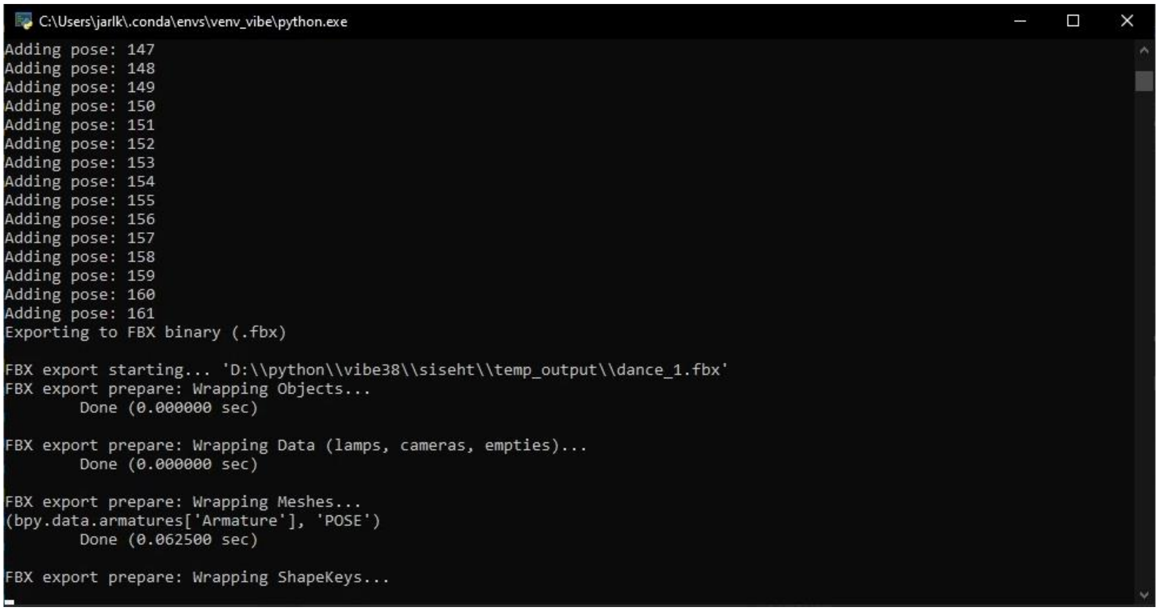
```
    **kw).encode(obj)
  File "C:\Users\jarlk\.conda\envs\venv_vibe\lib\json\encoder.py", line 199, in encode
    chunks = self.iterencode(o, _one_shot=True)
  File "C:\Users\jarlk\.conda\envs\venv_vibe\lib\json\encoder.py", line 257, in iterencode
    return _iterencode(o, 0)
  File "C:\Users\jarlk\.conda\envs\venv_vibe\lib\site-packages\django\core\serializers\json.py", line 105, in default
    return super().default(o)
  File "C:\Users\jarlk\.conda\envs\venv_vibe\lib\json\encoder.py", line 179, in default
    raise TypeError(f'Object of type {o.__class__.__name__} '
TypeError: Object of type IntegrityError is not JSON serializable
[19/May/2023 14:54:42] "POST /paths/upload/ HTTP/1.1" 500 212785
Running "ffmpeg -i D:\python\vibe38\siseht\temp_videos\dance.mp4 -f image2 -v error /tmp\dance_mp4/%06d.png"
Images saved to "/tmp\dance_mp4"
Input video number of frames 162
Running Multi-Person-Tracker
100%|████████████████████████████████████████████████████████████████████████████████| 54/54 [00:58<00:00,  1.01s/it]
Finished. Detection + Tracking FPS 2.79
WARNING: You are using a SMPL model, with only 10 shape coefficients.
WARNING: You are using a SMPL model, with only 10 shape coefficients.
=> loaded pretrained model from 'D:\python\vibe38\lib_package\data\vibe_data\spin_model_checkpoint.pth.tar'
Performance of pretrained model on 3DPW: 56.56075477600098
Loaded pretrained weights from "D:\python\vibe38\lib_package\data\vibe_data\vibe_model_wo_3dpw.pth.tar"
Running VIBE on each tracklet...
100%|██████████████████████████████████████████████████████████████████████████████████| 1/1 [00:15<00:00, 15.76s/it]
VIBE FPS: 10.28
Total time spent: 84.12 seconds (including model loading time).
Total FPS (including model loading time): 1.93.
Saving output results to "D:\python\vibe38\siseht\temp_vibe\dance\vibe_output.pkl".
================= END =================
```

Converter

The converter script imported the SMPL model asset to the Blender Python API and then applied the 3D data generated by the VIBE model to the SMPL model asset. It was then exported to a .fbx/.glb file. Figure 9 below shows the converter script running in the terminal.

Figure 8

Converter script running in the terminal



```
Adding pose: 147
Adding pose: 148
Adding pose: 149
Adding pose: 150
Adding pose: 151
Adding pose: 152
Adding pose: 153
Adding pose: 154
Adding pose: 155
Adding pose: 156
Adding pose: 157
Adding pose: 158
Adding pose: 159
Adding pose: 160
Adding pose: 161
Exporting to FBX binary (.fbx)

FBX export starting... 'D:\\python\\vibe38\\siseht\\temp_output\\dance_1.fbx'
FBX export prepare: Wrapping Objects...
        Done (0.000000 sec)

FBX export prepare: Wrapping Data (lamps, cameras, empties)...
        Done (0.000000 sec)

FBX export prepare: Wrapping Meshes...
(bpy.data.armatures['Armature'], 'POSE')
        Done (0.062500 sec)

FBX export prepare: Wrapping ShapeKeys...
```

Problems experienced by the Users

The responses provided by the interviewed individuals with varying levels of experience indicated that 3D animation was a commonly used tool in the industry. Some respondents worked with 3D animation on a regular basis, while others had minimal experience or were only somewhat familiar. Motion capture systems and video-to-3D animation converters were positively anticipated by some of the respondents to facilitate the animation process as to create the foundation of their projects in an efficient manner. However, using these tools came with several challenges, including the need to ensure the accuracy and cleanliness of motion data, and the requirement for manual adjustments to create natural movement and achieve a realistic result.

Acceptability of the System

The second objective of the study was to evaluate the level of the acceptability of the motion capture system in terms of improving the conversion of video files to 3D animation files as perceived by the users using the acceptability questionnaire.

Table 13

Descriptive Statistics of the acceptability questionnaire results

	N	Minimum	Maximum	Mean	Std. Deviation
Acceptability	50	4.00	5.00	4.8800	.25873
Valid N (listwise)	50				

Based on the descriptive statistics of the acceptability questionnaire results found in Table 13, the level of the acceptability of the motion capture system was very high (4.88 ± .26). The interpretation was taken from Table 6.

Usability of the System

The third objective of the study was to evaluate the usability of the motion capture system in terms of System Usefulness, Information Quality, Interface, and Overall Usability using the PSSUQ (Post-Study System Usability Questionnaire).

Table 14

Descriptive Statistics of the PSSUQ results

	N	Minimum	Maximum	Mean	Std. Deviation
System Usefulness	50	4.00	7.00	6.5100	.81121
Information Quality	50	4.00	7.00	6.6600	.62629
Interface	50	5.00	7.00	6.7200	.48613
Overall Usability	50	5.00	7.00	6.7000	.50508
Valid N (listwise)	50				

Based on the descriptive statistics of the PSSUQ results found in Table 14, the usability level of the motion capture system in terms of System Usefulness, Information Quality, Interface, and Overall Usability was very high (6.51 ± .81, 6.66 ± .63, 6.72 ± .49, 6.70 ± .51, respectively). The interpretation was taken from Table 7.

Efficiency of the Pose Detection Algorithm

The fourth objective of the study was to evaluate the level of efficiency of the pose detection algorithm. The pose detection algorithm utilized the VIBE model. Together with VIBE, the SPIN and Temporal HMR models were also evaluated to compare the results of each evaluation to find which of the models was more efficient.

The evaluation of the models was done on two commonly used 3D datasets, namely, the 3D Poses In The Wild Dataset, which has mostly data in outdoor conditions, using Inertial Measurement Unit (IMU) sensors to compute pose and shape ground truth (von Marcard et al., 2018), and the MPI-INF-3DHP dataset, which was captured in a multi-camera studio with ground truth from commercial marker-less motion capture (Mehta et al., 2017)

Table 15

Evaluation of Human Pose and Shape Estimation models on the 3DPW and MPI-INF-3DHP datasets based on the PA-MPJPE metric measured in millimeters

	Datasets	
Models	3DPW	MPI-INF-3DHP
Temporal HMR	76.7	89.8
SPIN	59.2	67.5
VIBE	56.5	63.4

Based on the evaluation found in Table 15, the VIBE model had lesser PA-MPJPE than the other models on almost all datasets, which can be interpreted as being more efficient than the others.

Figure 10

Qualitative comparison of video input frames to equivalent 3D animation frames.

(Sample 1)

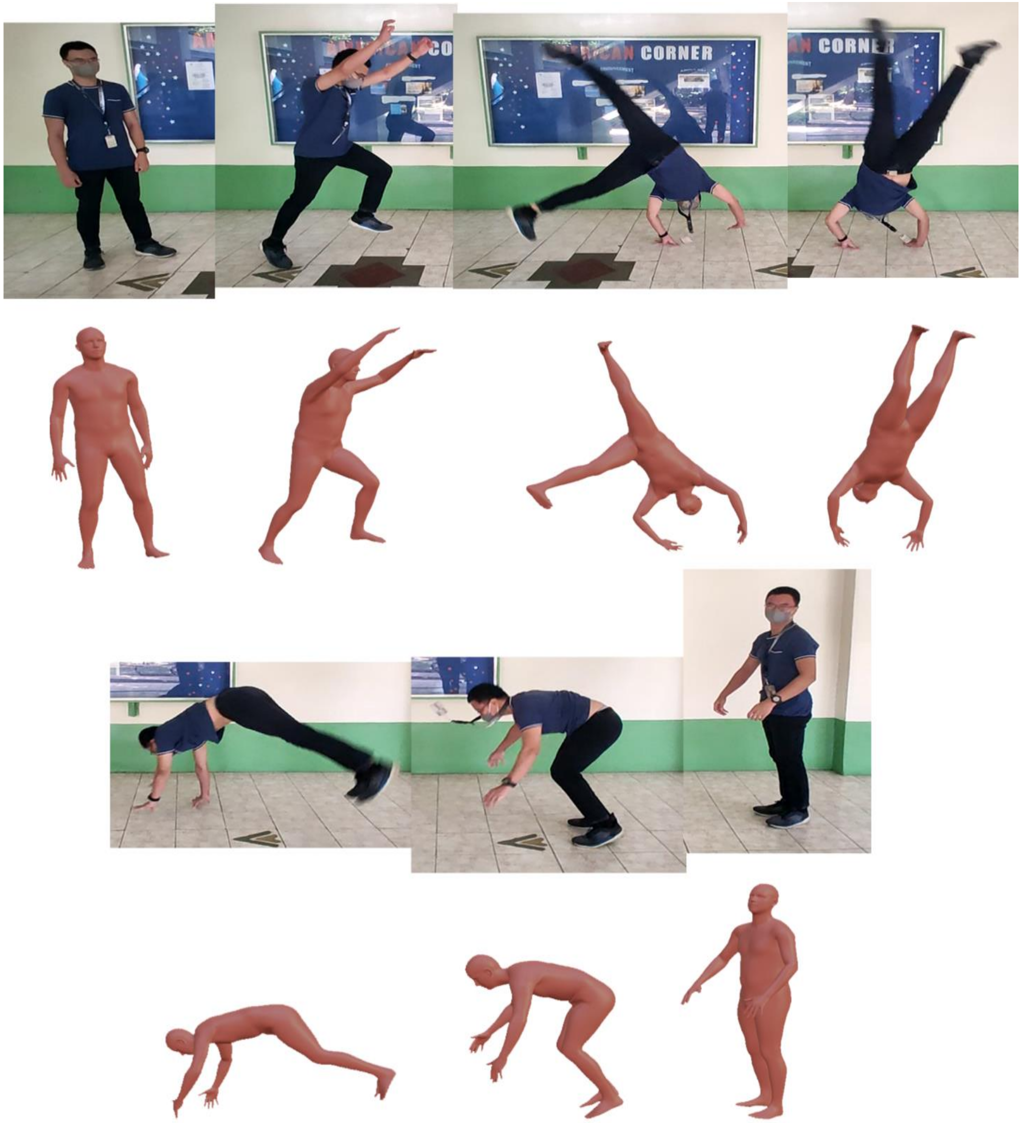


Figure 11

Qualitative comparison of video input frames to equivalent 3D animation frames.

(Sample 2)

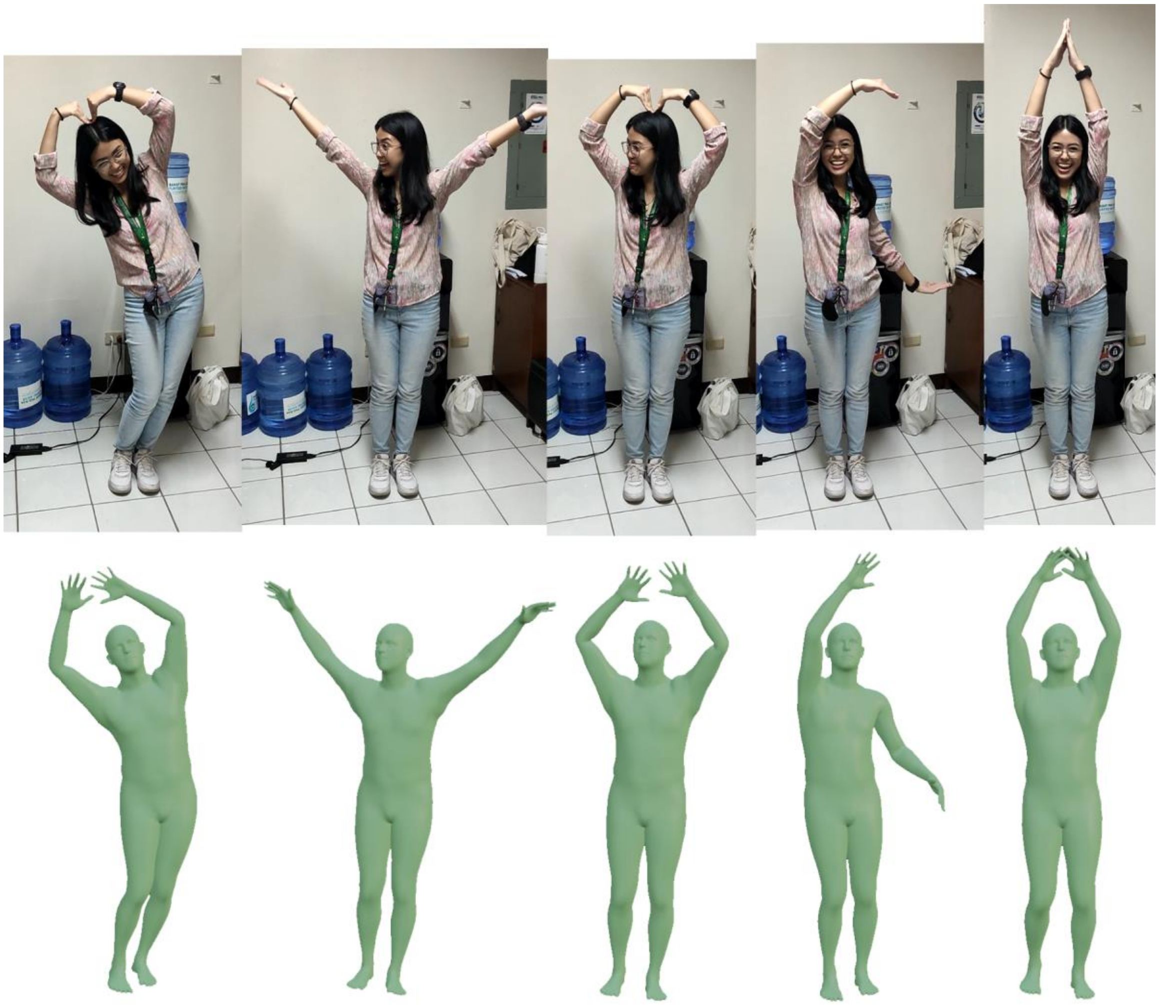
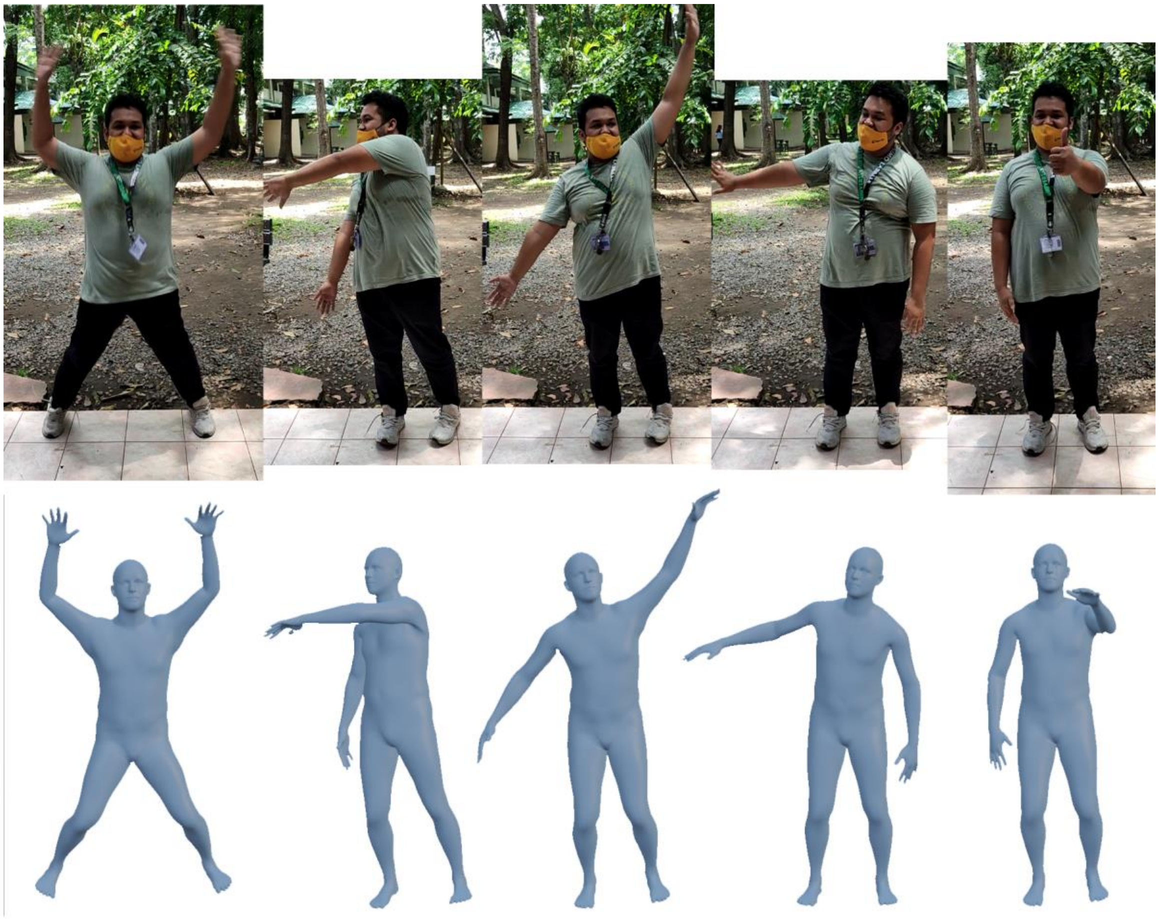


Figure 12

Qualitative comparison of video input frames to equivalent 3D animation frames.

(Sample 3)



Conclusions

Based on the results of the study, the following conclusions were formulated:

1. The design and development of the *motion capture system* with features that enable users to convert video files to 3D animation files and then download the generated 3D animation file addresses the problems experienced by users previously when converting video files to 3D animation files.
2. The *motion capture system* is very acceptable in terms of improving the conversion of video files to 3D animation files according to the users and participants of the study.
3. The *motion capture system* is also very usable in terms of System Usefulness, Information Quality, Interface, and Overall Usability.
4. The *VIBE pose detection algorithm* is more efficient compared to previous state-of-the-art models such as SPIN and Temporal HMR.

Implications

In terms of the acceptability of the system, many participants that tested the system expressed positive feedback on the functionalities and features of the system.

Verbal comments and survey tallies regarding the user interface of the website were also mostly positive with many positive remarks on the website accessibility and instructions that enabled them to easily use the system.

In terms of its overall usability and implication in the community, the system was easily used by the participants. Visual and interactive hindrances that prevented the usage of the system such as bugged UIs or unresponsive design were not experienced by the participants.

Finally, the system contributes on how motion capture can be utilized to simplify and ease the need of aspiring developers and animators in creating their own projects by only using a single camera and their preferred program without the requirement of expensive motion tracking equipment and green-screen studio and increase the public perception of how one can create motion capture animations without the aforementioned expensive requirements.

Recommendations

Based on the results of the study, conclusions, and implications made by the researchers, the following recommendations were formulated:

1. A multi-person pose detection can be implemented and utilized in order for more complex scenes involving multiple people to be converted.
2. Save and display the conversion history of the user so that they don’t need to convert the same files multiple times in certain circumstances such as the loss of converted files.
3. Implement hand landmark detection for further accurate hand animations and facial landmark detection for capturing of facial expressions.
4. Test the changing of the 3D model asset after converting to provide more customization for the users.

The challenges experienced by the users are important to consider for researchers and educators in the field of entertainment and multimedia computing, game development, and graphics in general. Continued research and education can help address these challenges and promote the development of more efficient and effective tools and techniques for 3D animation and motion capture.

REFERENCES

Akhter, I. (2015). *Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction*. https://openaccess.thecvf.com/content_cvpr_2015/html/Akhter_Pose-Conditioned_Joint_Angle_2015_CVPR_paper.html

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, Jitendra Malik.(2018, June 23). End-to-end Recovery of Human Shape and Pose. https://arxiv.org/pdf/1712.06584.pdf?fbclid=IwAR16jCnKgzaytqiljm9aTXTBa_oeo7B5u-3X7f2XXjnfCDPQ_6XYGabcNlM

Cao, C., Tang, Y., Huang, D., Gan, W., & Zhang, C. (2021). IIBE: An Improved Identity-Based Encryption Algorithm for WSN Security. *Security and Communication Networks*, *2021*, 1–8. https://doi.org/10.1155/2021/8527068

Catalin Ionescu, Dragos Papava, Vlad Olaru and Cristian Sminchisescu. (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, No. 7.* http://vision.imar.ro/human3.6m/pami-h36m.pdf

Catalin Ionescu, Fuxin Li and Cristian Sminchisescu. (2011). Latent Structured Models for Human Pose Estimation. *International Conference on Computer Vision.* http://vision.imar.ro/human3.6m/ils_iccv11.pdf

cbsudux (2019). Human Pose Estimation 101. https://github.com/cbsudux/Human-Pose-Estimation-101

Chen, C. (2017). *3D Human Pose Estimation = 2D Pose Estimation + Matching*. https://openaccess.thecvf.com/content_cvpr_2017/html/Chen_3D_Human_Pose_CVPR_2017_paper.html

Choi, J. (2019). Technology Trends for Motion Synthesis and Control of 3D Character. *Journal of the Korea Society of Computer and Information*, *24*(4), 19–26. https://doi.org/10.9708/jksci.2019.24.04.019

Datagen. https://datagen.tech/guides/data-training/ground-truth/

Elhayek, A., Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., & Theobalt, C. (2015, June). *Efficient ConvNet-based Marker-less Motion Capture in General Scenes with a Low Number of Cameras*. https://vcai.mpi-inf.mpg.de/projects/convNet_moCap/

Frankle, J. (2018, March 9). *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks*. arXiv.org. https://arxiv.org/abs/1803.03635

Gu, X. (2016a). *Dexmo: An inexpensive and lightweight mechanical exoskeleton for motion capture and force feedback in VR*. Dexmo: An Inexpensive and Lightweight Mechanical Exoskeleton for Motion Capture and Force Feedback in VR. https://www.repository.cam.ac.uk/handle/1810/256141

Han, F., & Bo, X. (2015). *Research and Literature Review on Developing Motion Capture System for Analyzing Athletes Action*. https://doi.org/10.2991/etmhs-15.2015.103

Howard, A. (2019, May 6). *Searching for MobileNetV3*. arXiv.org. https://arxiv.org/abs/1905.02244

Huang, G. (2017). *Densely Connected Convolutional Networks*. https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html

Iandola, F. N., Han, S., Moskewicz, M. W., & Keutzer, K. (2016a). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *ResearchGate*. https://www.researchgate.net/publication/301878495_SqueezeNet_AlexNet-level_accuracy_with_50x_fewer_parameters_and_05MB_model_size

Iskakov, K. (2019). *Learnable Triangulation of Human Pose*. https://openaccess.thecvf.com/content_ICCV_2019/html/Iskakov_Learnable_Triangulation_of_Human_Pose_ICCV_2019_paper.html

Kocabas, Muhammed and Athanasiou, Nikos and Black, Michael J. (2020, June). VIBE: Video Inference for Human Body Pose and Shape Estimation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/1912.05656

Loper, Matthew and Mahmood, Naureen and Romero, Javier and Pons-Moll, Gerard and Black, Michael J. (2015, October 6). SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*. https://smpl.is.tue.mpg.de/index.html

Mehta, Dushyant and Rhodin, Helge and Casas, Dan and Fua, Pascal and Sotnychenko, Oleksandr and Xu, Weipeng and Theobalt, Christian. (2017). Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. *3D Vision (3DV), 2017 Fifth International Conference on.* https://vcai.mpi-inf.mpg.de/3dhp-dataset/

Mező, K., & Mező, K. (2014). The IPOO-model of creative learning and the students' information processing characteristics. *Psihološka Obzorja*, *23*, 136–144. https://doi.org/10.20419/2014.23.414

Mo, C., Hu, K., Mei, S., Kuang, M., & Wang, Z. (2021). *Keyframe Extraction from Motion Capture Sequences with Graph based Deep Reinforcement Learning*. https://doi.org/10.1145/3474085.3475635

Moon, G. (2019a). *Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image*. https://openaccess.thecvf.com/content_ICCV_2019/html/Moon_Camera_Distance-Aware_Top-Down_Approach_for_3D_Multi-Person_Pose_Estimation_From_ICCV_2019_paper.html

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, Kostas Daniilidis. (2019, September 27). Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. https://arxiv.org/pdf/1909.12828.pdf?fbclid=IwAR3LnXn5C9tNU24KBFJ9Zv221aTfzSI1d6WopLvCn9GxydqdUZzPA0VYkBs

Okoyomon, A. (2019, December 12). *How Motion Capture Works - Science World*. Science World. https://www.scienceworld.ca/stories/how-motion-capture-works/

Pavlakos, G. (2018). *Ordinal Depth Supervision for 3D Human Pose Estimation*. https://openaccess.thecvf.com/content_cvpr_2018/html/Pavlakos_Ordinal_Depth_Supervision_CVPR_2018_paper.html

Pavllo, D. (2019). *3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training*. https://openaccess.thecvf.com/content_CVPR_2019/html/Pavllo_3D_Human_Pose_Estimation_in_Video_With_Temporal_Convolutions_and_CVPR_2019_paper.html

Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019a). Regularized Evolution for Image Classifier Architecture Search. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, *33*(01), 4780–4789. https://doi.org/10.1609/aaai.v33i01.33014780

Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. (2017, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8099626

Rhodin, H. (2018). *Learning Monocular 3D Human Pose Estimation From Multi-View Images*. https://openaccess.thecvf.com/content_cvpr_2018/html/Rhodin_Learning_Monocular_3D_CVPR_2018_paper.html

Rogez, G., Weinzaepfel, P., & Schmid, C. (2017a). *LCR-Net: Localization-Classification-Regression for Human Pose*. https://doi.org/10.1109/cvpr.2017.134

Rogez, G., Weinzaepfel, P., & Schmid, C. (2020). LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. https://doi.org/10.1109/tpami.2019.2892985

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. A. (2018). ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. IEEE Transactions on Intelligent Transportation Systems, 19(1), 263–272. https://doi.org/10.1109/tits.2017.2750080

Sandler, M. (2018, January 13). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. arXiv.org. https://arxiv.org/abs/1801.04381

Seethapathi, N. (2019, July 24). *Movement science needs different pose tracking algorithms*. arXiv.org. https://arxiv.org/abs/1907.10226

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). *Deep High-Resolution Representation Learning for Human Pose Estimation*. https://doi.org/10.1109/cvpr.2019.00584

Tong, Y., Weiran, C., Sun, Q., & Chen, D. (2021). The Use of Deep Learning and VR Technology in Film and Television Production From the Perspective of Audience Psychology. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.634993

Trumble, M., Gilbert, A., Madison, C., Hilton, A., & Collomosse, J. (2017). *Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors*. https://cvssp.org/projects/totalcapture/TotalCapture/

von Marcard, T., Pons-Moll, G., & Rosenhahn, B. (2016). Human Pose Estimation from Video and IMUs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(8), 1533–1547. https://doi.org/10.1109/tpami.2016.2522398

von Marcard, Timo and Henschel, Roberto and Black, Michael and Rosenhahn, Bodo and Pons-Moll, Gerard. (2018). Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. *European Conference on Computer Vision (ECCV)*. https://virtualhumans.mpi-inf.mpg.de/3DPW/

Wei, Y. (2022). Deep-Learning-Based Motion Capture Technology in Film and Television Animation Production. *Security and Communication Networks*, *2022*, 1–9. https://doi.org/10.1155/2022/6040371

Williams, J. (n.d.). *Recovering Accurate {3D} Human Pose in The Wild Using {IMUs} and a Moving Camera | Perceiving Systems - Max Planck Institute for Intelligent Systems*. Max Planck Institute for Intelligent Systems. https://ps.is.mpg.de/publications/vip-eccv-2018

Xiang, D. (2019). *Monocular Total Capture: Posing Face, Body, and Hands in the Wild*. https://openaccess.thecvf.com/content_CVPR_2019/html/Xiang_Monocular_Total_Capture_Posing_Face_Body_and_Hands_in_the_CVPR_2019_paper.html

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. https://doi.org/10.1109/cvpr.2017.634

Xu, W. (2017a, August 7). *MonoPerfCap: Human Performance Capture from Monocular Video*. arXiv.org. https://arxiv.org/abs/1708.02136

Zanfir, A. (2018a). *Deep network for the integrated 3D sensing of multiple people in natural images*. Lund University. https://portal.research.lu.se/en/publications/deep-network-for-the-integrated-3d-sensing-of-multiple-people-in-

Zeng, R. (2021). Research on the Application of Computer Digital Animation Technology in Film and Television. *Journal of Physics*, *1915*(3), 032047. https://doi.org/10.1088/1742-6596/1915/3/032047

Zhang, X. (2018). *ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices*. https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_ShuffleNet_An_Extremely_CVPR_2018_paper.html

Zhou, D., Xinzhu, F., Yi, P., Yang, X., Zhang, Q., Wei, X., & Yang, D. (2019). 3D Human Motion Synthesis Based on Convolutional Neural Network. *IEEE Access*, *7*, 66325–66335. https://doi.org/10.1109/access.2019.2917609

Appendix A

USABILITY QUESTIONNAIRE

This survey is being conducted by Group 1 as part of the requirements for Computer Science in Thesis 1 at the University of St. La Salle.

The purpose of this questionnaire is to evaluate the motion capture system in terms of its Overall Usability, System Usefulness, Information Quality, and Interface Quality using the PSSUQ (Post Study System Usability Questionnaire) standard as the basis.

Please do not leave any items unanswered. Also, if you agree to respond, your answer will remain strictly confidential neither your name nor your individual responses will not be given to any other individual or organization either inside or outside of the University of St. La Salle.

Your cooperation and honesty are greatly appreciated.

Part I - Profile of Respondent

Name (Optional): ___________________________________ Date:________________

Organization/ Agency affiliated: ____

Direction: Please rate the system corresponding to your satisfaction by putting check (✓) for every criterion opposite to it.

No.		Strongly Disagree				Strongly Agree			
		1	2	3	4	5	6	7	N/A
1	Overall, I am satisfied with how easy it is to use this system.								
2	It was simple to use this system								
3	I was able to complete the tasks and scenarios quickly using this system.								

4	I felt comfortable using this system.								
5	It was easy to learn to use this system.								
6	I believe I could become productive quickly using this system.								
7	The system gave error messages that clearly told me how to fix problems.								
8	Whenever I made a mistake using the system, I could recover easily and quickly.								
9	The information(such as online help. on-screen messages, and other documentation) provided with the system was clear.								
10	It was easy for me to find the information I needed.								
11	The information was effective in helping me complete the tasks and scenarios								
12	The organization of information on the system screens was clear.								

13	The interface of this system was pleasant.								
14	I liked using the interface of this system.								
15	This system has all the functions and capabilities I expect it to have.								
16	Overall, I am satisfied with the system.								

Comments:

____

____

____

Signature of the Evaluator

Appendix B

ACCEPTABILITY QUESTIONNAIRE

This survey is being conducted by Group 1 as part of the requirements for Computer Science in Thesis 2 at the University of St. La Salle. The purpose of this questionnaire is to evaluate the motion capture system in terms of its acceptability.

Please do not leave any items unanswered. Also, if you agree to respond, your answer will remain strictly confidential neither your name nor your individual responses will not be given to any other individual or organization either inside or outside of the University of St. La Salle. Your cooperation and honesty are greatly appreciated.

Direction: Please rate the system by checking the box with the following scale:

1 - Strongly Disagree, **2** - Disagree, **3** - Neutral, **4** - Agree, **5** - Strongly Agree

Criteria for Acceptability	1	2	3	4	5
1. I am satisfied with the whole experience in converting video to 3D animation with the motion capture system.					
2. I found it easy to learn how to use the motion capture system in converting the file to 3D animation.					
3. I found it easy to use the motion capture system in converting the file to 3D animation.					
4. I enjoyed using the motion capture system in converting the file to 3D animation.					
5. I was able to upload the file easily without encountering any problems in converting the file to 3D animation.					
6. I was able to download the file without any problems in converting the file to 3D animation.					
7. I am satisfied with the output of the motion capture conversion.					
8. Overall, it was acceptable to use the motion capture system.					

__________________________ ________________

Signature over printed name Date

Appendix C

VALIDATION OF SURVEY INSTRUMENT

Using the criteria developed for evaluating survey questionnaires set forth by Carter V. Good and Douglas B. Scates, please evaluate the attached self-made survey instrument for the proposed study by checking the respective boxes provided herein.

Rating Scale:	5 – Strongly Agree	3 – Undecided	1 – Strongly Disagree
	4 – Agree	2 – Disagree	

Criteria for Validity	1	2	3	4	5
1. This questionnaire is short and respondents expect it would not drain much of their precious time.					
2. The questionnaire is interesting and has a face appeal such that respondents will be induced to respond to it and accomplish it fully.					
3. The questionnaire can obtain some depth from respondents and avoid superficial answers.					
4. The items/questions and their alternative responses are not too suggestive and not too stimulating.					
5. The questionnaire can elicit responses which are definite but not mechanically forced.					
6. Questions/items are stated in such a way that the responses will not be embarrassing to the person/persons concerned.					
7. Questions/items are formed in such a manner as to avoid suspicion on the part of the respondents.					
8. The questionnaire is not too narrow nor restrictive nor limited in its philosophy.					
9. The reasons for the questionnaire when taken as a whole could answer the basic purpose for which the questionnaire is designed and therefore, considered valid.					

__________________________ ________________

Signature over printed name Date

Appendix D

INFORMED CONSENT

The survey is being conducted by Karr Christopher Balbin, Kean Gabriel Canaria, Kent Matthew Estiamba, and Jarl Keenen Sarmiento as part of the requirements for the degree of BS Computer Science under the course Thesis 1 at the University of St. La Salle.

The title of the study is 3D MOTION CAPTURED ANIMATION USING POSE DETECTION ALGORITHMS. The goal(s) of the study is to develop a tool that converts video recordings to a 3D animation file.

The purpose of this questionnaire is to evaluate the usability of the motion capture system in terms of its Overall Usability, System Simplicity, Information Quality, and Interface Quality using PSSUQ (Post Study System Usability Questionnaire) standard as bases for the evaluation of the software quality.

Your participation is voluntary and there will be no penalty should you not agree to participate. You also have the right to withdraw your participation at any time. If you agree to respond, your answers will remain strictly confidential. Neither your name nor your individual responses will be given to any other individual or organization either inside or outside of the University of St. La Salle Bacolod.

Should you have questions or clarification regarding the research, you may contact the following researchers: Karr Christopher Balbin at 09763120766 or balbinkc@gmail.com, Kent Matthew Estiamba at 092129790765 or kentestiamba@gmail.com and Jarl Keenen Sarmiento at 09206151240 or jarlkeenen@gmail.com.

If you have any concerns about your rights or treatment as a research participant, you may contact the CET Ethics Chairperson, Engr. Rendell Barcimo at 4346100 loc 132 or at r.barcimo@usls.edu.ph.

Your cooperation and honesty are greatly appreciated.

____

Signature over printed name of participant

____

Signature over printed name of witness

Appendix E

Curriculum Vitae

Personal Information

Name	: Jarl Keenen A. Sarmiento
Address	: Blk. 10, Lot 7, Pygros Rd., Santorini Subd., Brgy Madalagan, Bacolod City, Negros Occidental
Contact No.	: 09206151240
Email	: jarlksarmiento@gmail.com
Nationality	: Filipino
Date of Birth	: February 1, 2001

Education

- **Bachelor of Science in Computer Science**

  University of St La Salle Bacolod

Personal Information

Name : Kent Matthew V. Estiamba

Address : Blk. 8, Lot 21, Raymund Street.,
Brgy Singcang-Airport, Bacolod City, Negros Occidental

Contact No. : 09212979065

Email : kentestiamba@gmail.com

Nationality : Filipino

Data of Birth : December 17, 2000

Education

- **Bachelor of Science in Computer Science**

  University of St La Salle Bacolod

Personal Information

Name	:	Karr Christopher G. Balbin
Address	:	Lot 55, Aries Street, Sharina Heights, Brgy. Taculing, Bacolod City, Negros Occidental, Philippines
Contact No.	:	0976 312 0766
Email	:	balbinkc@gmail.com
Nationality	:	Filipino
Data of Birth	:	February 22, 2001

Education

- **Bachelor of Science in Computer Science**

  University of St La Salle Bacolod

Personal Information

Name : Kean Gabriel F. Canaria

Address : Blk 8, Lot 11, St. Paul’s Village, Talisay City, Negros
Occidental, Philippines
Contact No. : 09453014199

Email : Keancanaria@gmail.com

Nationality : Filipino

Data of Birth : March 2, 2001

Education

- **Bachelor of Science in Computer Science**

  University of St La Salle Bacolod