AALBORG UNIVERSITY
DENMARK

Hybrid vision-based motion capture for a human leg

Miguel Gargallo Vázquez

AALBORG UNIVERSITY
STUDENT REPORT

4th Semester
Vision, Graphic and Interactive Systems
Fredrik Bajers vej 7C
9220 Aalborg Ø
http://www.sict.aau.dk

Title:
Hybrid vision-based motion capture for a human leg

Project Theme:
Computer Vision

Project period:
February 2017–September 2017

Miguel Gargallo Vázquez

Supervisors:
Thomas B. Moeslund

Completed: 2017-09-08

Abstract:

The scope of this project is: computer vision and computer graphics, where the focus is on motion capture.
Marker-based and marker-less motion capture systems present a opposing set of advantages and disadvantages in terms of convenience and accuracy that could be made use of by combining them in a hybrid system.
The aim of the project is to provide the necessary background knowledge and a methodology to find a combination of both systems that would result in a more convenient set up.
An independent implementation and tests are carried out in a simplified scenario where a human leg moves and walks along a line. It is recorded with a camera and a marker-based system, using a set of markers attached to the different segments of the leg. The segments obtained in both systems are combined in different ways and the angles for the knee and ankle are computed and compared to those from the full marker set.

Contents

1 Problem Analysis

1.1 Project Proposal

The idea for this project comes from the opposite strengths and limitations that marker-based and marker-less systems present: marker-based systems can be highly accurate but are inconvenient in terms of setup and flexibility, whereas current marker-less systems are not accurate enough for many applications but are extremely convenient.

Different ways of combining these systems will be explored, discussing how they help overcome the weaknesses of both systems, as well as the benefits and drawbacks of the chosen approach.

It would be interesting to understand how marker-based and marker-less systems could be combined and what are the results. These results would come in the way of correlation levels in comparison to the original high-accuracy marker-based system, by extracting several biomechanical parameters from the movement of the estimated segments.

Said scenario will be explored in this work along with a much more simple one, consisting of a single leg being tracked while moving in a plane. This scenario keeps the same goals but at a much smaller scale, allowing to make assumptions on the nature of the data, and therefore resulting in a more straightforward development of a test.

The potential of this experiment lies in the fact that, when the marker-less system provides enough accuracy for a given segment, said segment could be dropped from the marker-based system. This means a smaller set of markers would be needed, resulting in reduced set up requirements. Furthermore, it allows a user to decide a trade-off between convenience and accuracy.

The goal of this report is to document the work done, both the ad-hoc solution for the simplified experiment and the different alternatives for a realistic situation; and ultimately to prove that marker-based and marker-less systems can be combined to achieve halfway results.

1.2 Background Knowledge

This section will describe some of the basic theory behind the systems, to better understand how they work and, most importantly, to understand the challenges of combining them. It will not discuss the theory behind the biomechanical estimations or models, as it is out of the scope of this project.

1.2.1 Motion capture systems

The topic of this project is motion capture systems. These are computer systems that use a variety of algorithms to estimate the motion of a subject throughout a series of video inputs.

They have met extensive use in the film and video game industries, where the movements of an actor are recorded and later translated into a virtual character. In these scenarios, an animator must usually clean most of the result of small errors, usually tuning it to better fit the intended artistic tone of the project; nonetheless, many hours of work are saved by having the bulk animation done through motion capture.

Other common areas of use for these systems are research in biomechanics, sports analysis, and rehabilitation. In this case, the interest lies in obtaining objective data about the movement of specific parts of the subject.

High accuracy is of course always a desirable feature, but it is important to clarify how its importance varies between the different cases. The significance of inaccuracies in the data is much smaller in an animation, where an animator can fix any errors but it is important to be able to record quickly, compared to a project in biomechanical research where accurate numerical data is essential. This creates a spectrum of needs in terms of the trade-off between accuracy and convenience that different systems aim to satisfy.

Surveys from different areas that employ human tracking systems, such as Augmented Reality[1], Virtual Reality[2] and rehabilitation medicine[3], make the common distinction between visual (or videometric), non-visual and hybrid systems.

Visual systems employ different types of cameras to determine the position of visual landmarks associated with body parts, either directly from the shape of the body or through the use of additional attachments—typically referred to as markers—that are easily identifiable. These two kinds of visual tracking systems are commonly referred to as marker-based and marker-less motion capture, and the method of choice influences the type of camera used, as will be explained in more detail shortly.

In a typical vision-based motion capture scenario, the subject performs movement inside of a designated scene, with one or more cameras having somewhat unoccluded vision of the subject. Usually, the subject is required to wear special clothing in the form of tight and sometimes dark clothes that cover them as much as possible. Of course, the exception exists when the clothing is relevant to the performance, such as a dancer performance dressed in a gown.

Non-visual systems use other kinds of sensors, as opposed to light sensors (such as cameras). As summarised by Zhou et al.[3]: "Sensors employed within these systems adhere to the human body in order to collect movement information. These sensors are commonly categorised as mechanical, inertial, acoustic, radio, or microwave and magnetic based.". A possible addition to this list is non-visual sensors placed on the scene where the tracking takes place, instead of the body, such as force plates. Baillot et al.[2] review in depth how these tracking systems work, as well as some of their strengths and weaknesses. It

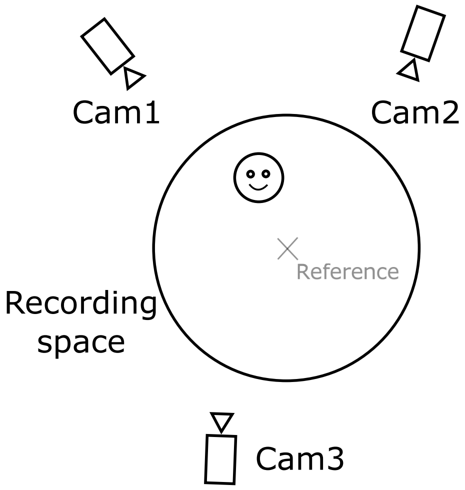

Figure 1.1: *Typical motion capture scene*

is important to notice that each of these present different advantages and disadvantages, which determines when they are to be used.

Hybrid systems combine different tracking systems among the ones mentioned previously, in order to work around some of their weaknesses. One common approach is to combine a visual, marker-less system, which operates with high frequency and does not interfere with the movements of the subject, with a more involved method that provides more stability and is less susceptible to noise. Van Krevelen et al.[1] refers to hybrid approaches as "the most promising way to deal with the difficulties posed by general indoor and outdoor mobile AR environments", and Zhou et al.[3] notes "although still at an experimental stage, have already demonstrated promising performance".

There is therefore a growing interest in combining existing techniques to create a more appropriate solution for the needs of the application. The array of sensors that can be combined make this approach quite versatile, although different combinations ask for different methodologies and requirements.

Vision-based human tracking has experienced great interest within computer vision research in recent decades. Several well-known surveys follow the advances in such research: Moeslund et al. review the most prominent work between 1980–2000[4] and 2000–2006[5] in marker-less approaches, Yang et al.[6] explore different aspects of current marker-less approaches and compare them from a biomechanical standpoint.

These improvements in the past decade make it worth looking into additional hybrid solutions involving visual marker-less tracking.

A study carried out in the motion-capture company Simi[7] by Becker et al.[8] shows promising results using their marker-less silhouette-based tracking system (Simi Shape 3D[9]), and their marker-based tracking system (Simi Motion 3D[10]), greatly improving the results obtained from the marker-less system alone.

As can be seen in table 1.2.1, correlation levels improve globally by adding a small subset of

joint	movement	marker-less correlation	hybrid correlation
hip	flexion/extension	0.82 (±0.21)	0.99 (±0.02)
	abduction/adduction	0.51 (±0.45)	0.94 (±0.09)
	rotation	0.41 (±0.32)	0.93 (±0.05)
knee	flexion/extension	0.98 (±0.03)	
ankle	plantar/dorsal flexion	0.91 (±0.08)	0.96 (±0.04)
	eversion/inversion	0.26 (±0.42)	0.38 (±0.33)
	abduction/adduction	0.34 (±0.40)	0.40 (±0.45)
shoulder	flexion/extension	0.96 (±0.02)	
	abduction/adduction	0.89 (±0.16)	0.94 (±0.05)
	rotation	0.49 (±0.50)	0.95 (±0.03)
elbow	flexion/extension	0.42 (±0.70)	0.92 (±0.07)

Table 1.1: *Correlation in a pure marker-less and a hybrid approach compared to a full markerset; extracted from Becker et al.[8]*

markers, in some cases to great extent. It is important to notice how different movements exhibit very different correlation levels, in part due to how they influence the silhouette of the body differently. It can be seen from the data that the marker-less system can track some movements very accurately on its own, which further motivates this work to explore hybrid systems: for many applications, the necessary marker set to meet some accuracy goals will be quite small.

However, both of the systems used by Becker et al.[8] are commercial solutions involving complex skeleton models focused on human tracking, and the dataset gathered is specific to the experiment, making it difficult to compare it to other existing solutions. Yang et al.[6] describe how most of the work in marker-less motion capture research uses specific datasets that are rarely shared for privacy purposes.

Instead, Yang et al. argue in favor of using publicly available datasets, of which the most common is HumanEva, gathered by Sigal et al.[11]. Such dataset is however extremely challenging, and making use of it is out of the scope of this project due to time constraints.

There are not many additional sources involving hybrid solutions combining marker-less and marker-based visual tracking. This project will look further into this approach from a more simplistic scenario that does not involve a commercial solution. Firstly, both tracking methods will be discussed.

1.2.2 Marker-based motion capture

Marker-based motion capture uses attachments to the body of the subject to estimate the configuration of the body, usually by starting at a known position where the markers can be identified, and interpreting the movement to the next frame.

Several types of markers exist, although they all share the purpose of being easily extracted from an image by acting as visual landmarks—typically appearing as bright points, although some areas such as Augmented Reality might make use of fiducial markers.

- Active markers emit a bright light, such as a LED.
- Passive markers use highly reflective material on their surface, and must be illuminated by an external light source. It is common to attach said source to the camera to ensure that all markers visible by the camera are properly illuminated.

It is common to employ infrared lighting to avoid confusing other light sources as markers: infrared LEDs in an active marker, or infrared light sources and reflective material in a passive system.

Once the markers have been extracted from an image, they must be uniquely identified. This is essential in order to understand the configuration of the body of the subject. In both cases, a known position can be used at the beginning of the recording, from which the identification process can be manually or automatically done in a more trivial way. From there, each consecutive frame will use information from the previous frame to estimate the new position of each of the markers, which becomes a tracking problem. Additionally, active markers can emit light in different patterns so they can be more easily identified.

The process of estimating the configuration of the body from the identified markers varies depending on the subject. In the case of a human subject, the process typically involves estimating the joint centres for the different segments from the marker positions. This is a biomechanics problem and will not be discussed in this project. A skeleton model can be built from these joint centres; the relevance of the choice of skeleton model will be later discussed.

Besides the already discussed requirements in term of setup, marker-based systems present a series of challenges:

- Occlusion. As the subject moves, self-occlusion occurs. Many markers will be placed on the opposite side of the subject from a camera's perspective; additionally, the subject's limbs can occlude markers on their body. This is solved to some extent by using several cameras around the subject. However, there still exist cases where a marker is occluded to all cameras, or where the amount of cameras that see it is so low that the data is ambiguous. Several techniques exist for finding lost markers, although manual interaction might be necessary if they don't succeed.
- Setup. As has already been mentioned, marker-based systems require a lengthy setup step due to the placement of the markers on the subject. This must often been done by an especialist and takes a considerable amount of time.
- Joint centre estimation. A regression equation is used to estimate the centre of the joints from marker positions. These equations are subject to some error, depending on the marker set and the equation. In the work of Sandau et al.[12], MRI scans are used to establish a quantitative comparison between different joint centre estimation methods, showing significant prediction errors and proposing new equations that minimize these. As is mentioned previously and in their work, it is common in vision-based motion capture to take marker-based systems as ground truth, even though they are not.

1.2.3 Marker-less motion capture

Marker-less systems do not use any attachments on the body of the subject. The system captures the silhouette of the subject as well as other features depending on the hardware and approach used. Colour cameras and depth cameras are common choices.

There are two distinct approaches in marker-less motion capture:

- Model-based: it uses an internal representation of the subject and how they can move. In the case of a human subject, it may use a simplified skeleton that is uses to match against the current frame, obtaining a skeleton configuration as a result. The matching process will be explained shortly. The main disadvantage is the need for an accurate model of the subject, which encumbers the convenience of the system.
- Model-free: it estimates the shape of the subject without additional information about it. It is agnostic to the identity of the subject, which gives this approach its main strength. On the other hand, the output of this approach consists of the shape of the object, but has no information about its configuration, which limits the number of applications that can make use of this approach. Other implementations using this approach employ additional information from the subject, such as the one from Dou et al.[14], obtain high fidelity reconstructions of subjects with complex clothing and movements utilizing previously captured 3D scans of the subject. This yields incredible visual fidelity at the expense of convenience and flexibility.

Yang et al. conclude that model-based systems provide smaller error than model-free, with Corazza et al.[15] being the best performing system. No implementation of this method was available at the time of writing, so it was discarded as an option.

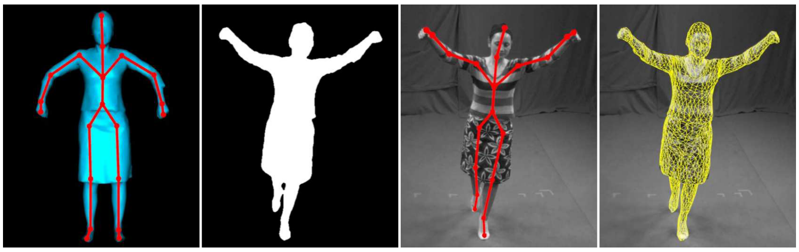

Figure 1.2: *Model-based pose estimation by Starck and Hilton[13].*

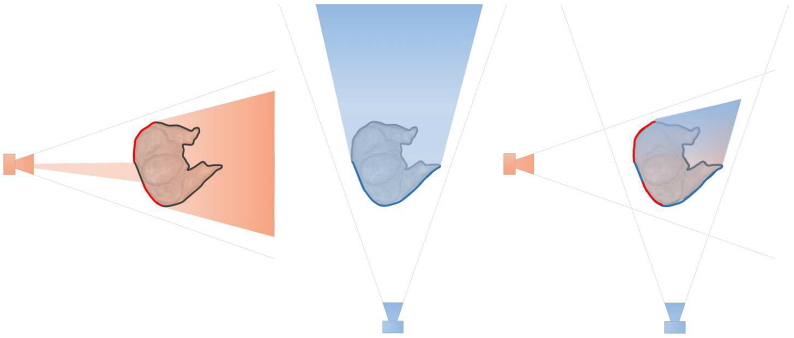

Figure 1.3: *Model-free performance reconstruction by Dou et al.[14].*

The steps taken during the process of a model-based system change a great deal between different lines of research; but incremental methods using image features are common, where the estimated pose from one frame is used in the next frame.

Additionally, image segmentation is a step necessary before any features can be extracted from the subject in the images. A common method usually available in Computer Vision tools (such as OpenCV) is Mixture of Gaussians (MOG), where each pixel is modeled by a mixture of Gaussians, as suggested by the name. However, the choice of the segmentation method is highly determined by the environment in which the recordings are taken.

Regarding incremental methods, e.g. a Kalman filter, the accuracy of the first frame is crucial. It is therefore common to set strong constraints on said first estimation, such as expecting the subject to assume a T-pose looking toward a specific camera, where the palms of the hands point down, the knees and head point forward, the elbows point back, and the rotation on the hips is minimal. Another set of assumptions is possible, of course.

One example of such incremental methods is the work from Starck and Hilton[13]. It uses several viewpoints, with the errors obtained from each viewpoint are added together to form a final error calculation that is used for the calculations. Three fundamental steps are described for each frame:

1. Local Optimization. Texture and contour features are extracted from all viewpoints and matched to those from the previous frame, obtaining a reference of the movement between both frames. One by one, the joints in the model are modified to align best with the features. This is a very efficient procedure, and allows different branches in the skeleton tree to be computed separately, but it also cannot recover easily from errors; moreover, errors are carried out through a branch, meaning that an error in a shoulder estimation will be carried out to the elbow and wrist estimations.
2. Global Optimization. In order to solve these errors that might affect entire chains from the Local Optimization step, this second and more fine-tuning step compares the contours of the projected model mesh and the image, measured with pixel precision. Also, the predicted pose is used as a stabilization method. This is a very slow process, but it will only be performed in those joints that need fixing. This is measured by an error threshold after Local Optimization that labels joints as misaligned to fix those joints' entire chains.
3. Surface Estimation. This step is meant to account for loose clothing and artifacts due to the way flesh behaves. It does so by dropping the constraint that originally ties skeleton model and 3D mesh together to handle geometry changes that aren't related directly to the segment orientations.

Just like in marker-based systems, occlusion in marker-less systems is a challenge that is partially solved by placing several cameras around the subject, when possible. Another shared challenge is the effects of clothing and loose skin, which affects marker-less in the same way.

This project focuses on motion capture on humans. One of the problems when comparing different research on marker-less motion capture is the absence of golden standard to compare against. This is discussed in depth in the review done by Yang et al.[6]. A modern marker-based system is usually chosen as ground truth, but as was mentioned before, this is not without error. However, at the current state of accuracy in marker-less, it can be considered a fair comparison, since the error is still orders of magnitude higher. Another problem is the lack of a common dataset between different research projects. HumanEva is the most commonly used dataset, although quite challenging. However, most research uses its own dataset, which is commonly not public to protect the privacy of the test subjects. On top of that, different measurements are used for quantifying error. All of this makes it impossible to properly compare these methods.

1.3 Problem Description

This project focuses on motion capture, and some of the shortcomings of typical motion capture systems. Marker-based systems, which use markers attached to the body of the subject, are highly accurate but inconvenient, due to the extensive setup they require. Marker-less systems, which use colour or depth cameras to estimate the shape of the body of the subject, are convenient in terms of setup but are not as accurate.

The accuracy of marker-less systems makes it usable for scenarios such as the gaming and film industry, where an animator can fix any errors, but unusable for biomechanical research or sports analysis. On the other hand, the cumbersome setup of marker-based systems limits the number of applications they can be used for as well.

Examples of the limitation of marker-based systems due to its setup have been documented in research as well, where a less accurate marker-less system gives better results due to the bigger amount of data they are able to capture[16].

This can be generalised to cases where machine learning is applied; the ease of recording with a marker-less system might make up for its lower accuracy due to the bigger dataset that can be created with the same effort. This difference will likely become more apparent as the accuracy for marker-less catches up to that of marker-based.

The way these opposing advantages and disadvantages line up motivate this work to investigate a hybrid approach, following big improvements in the recent years in marker-less motion capture and preliminary tests such as the one at the motion capture company Simi by Becker et al[8].

The project can also be seen as an enabling technology that would allow researchers to perform user studies with a less cumbersome set up, maximizing the trade-off they need between convenience and accuracy.

Through experiments carried out by students from *Aalborg University* at *Qualisys* in Gothenburg, it was noted that the marker-less system used has varying correlation values for the different body segments being tracked. Some of these segments consistently yielded satisfying results, and were deemed usable by a specialist for a sports analysis application, whereas other segments were results too inaccurate to be of any use. This agrees with the results reported at Simi[8].

It is expected that a project using this idea could be of great interest for all these applications of motion capture where using marker-based system is possible, and where the trade-off between convenience and accuracy is interesting, as opposed to a situation where accuracy should be as high as possible throughout the entire subject. The final condition is that some current marker-less system provides results that are close to acceptable but not quite there yet.

1.4 Delimitations

There is a series of limitations imposed to the project due to time constraints or being out of scope for the learning goal:

Highest accuracy This project will not try to deal with how to combine data coming from both systems to obtain better accuracy than the original marker-based system. It instead focuses on reducing the necessary marker set to carry out a given application.

Disagreements It will not explore what decisions should be taken when both systems disagree, e.g.in order for a segment from the marker-less to be incorporated into the marker-based skeleton, the previous or following segment in the marker-based skeleton would have to be moved since it makes no sense anatomically.

Marker dependency Markers should affect as few joints as possible. This is discussed more in depth in Design.

2 Design

This work explores a solution that could have an impact on scenarios where a marker-less motion capture system gives satisfactory results for some body segments but unsatisfactory results for others.

Extensive research exists for recovering from error when markers go missing due to occlusion. Classically this is done by interpolating the data between the points where the marker is lost and found, but the longer time the marker is lost, the more complex the movement might be. Some methods incorporate biomechanical data and use other markers as reference as to where the missing marker could and could not have been.

Using marker-less data could provide a strong reference to keep the missing marker virtually restricted to a specific area, since missing a marker for a specific segment does not necessarily mean that the segment will be missing from the marker-less system. In this sense, the system recovers from error in a more straightforward way.

In a similar way, the segment with a missing marker could be dropped from the marker-based system for the duration of these frames and the segment data could be obtained simply from the marker-less system. This could be seen as an alternative application where the marker-less data is only used on demand for specific segments and moments and remains "silent" for the rest of the performance.

The opposite scenario is also possible, where only a handful of markers are attatched to establish boundaries in the marker-less system, as explored by Becker et al.[8].

As a proof of concept, a smaller experiment was carried on, where challenges unrelated to the subject could be removed for the sake of simplicity.

To prove that halfway results are possible by combining marker-based and marker-less systems, one single body segment was recorded with both systems and then both data used in different combinations as part of several tests. These tests will be explained in detail later on.

The recording consists of a single leg moving roughly along one plane, showing different flexing motions and walking. The data to be extracted is the orientation of the different segments within the leg, namely: thigh, shank and foot; or from the perspective of the joints, the rotation of knee and the ankle.

2.1 Marker-based setup

For availability reasons, the setup consists of a set of HTC Vive trackers instead of the more common reflective or emissive markers discussed earlier. These trackers determine their own 3D location by using a set of special light emitting sources as reference. These sources are commonly referred to as lighthouses due to the way they emit light, in a very similar fashion to a lighthouse: a light source is surrounded by a spinning rotor that has a small hole. Two of these rotors are used in each lighthouse, one spinning vertically and one spinning horizontally.

The process begins with the lighthouses and the trackers being synchronised, which, in the version of the HTC Vive used, is done by using a bright light and starting a counter simultaneously on all the devices when said light is picked up by the trackers. Once the devices are in sync, the rotors start spinning. The time it takes for a tracker to observe the light through the spinning rotor can be directly translated to the angle between the lighthouse and the tracker. By using a vertical and a horizontal rotor, the tracker can determine both the horizontal and vertical angles, allowing it to calculate its own position. This position is a 3D point, which makes it behave similarly in terms of output to a more conventional marker-based system.

The setup consists of six trackers, two of which are actually controllers and can process additional input from the user—but this is irrelevant for the experiment, since they track their position in the same way. They are attached to the leg as tightly as possible to avoid noise in the data, and are distributed in pairs for each leg segment; this makes it straightforward to compute the rotation angle from the lines created by each pair of trackers. figure 2.1 shows the positions of the trackers.

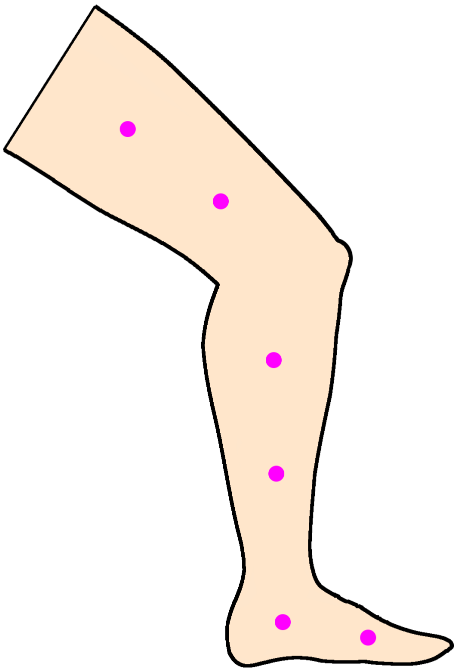

Figure 2.1: *Position of the six trackers on the leg*

2.2 Marker-less setup

Since the leg will be moving mostly along a plane, and therefore roughly in only two dimensions, a camera can be placed in a parallel plane at an appropriate distance that will record the necessary range. This way, the whole movement can be recorded with a single camera. The recordings will consist of colour video from this single viewpoint.

The recognition module encompasses two steps: foreground detection and segment recognition.

Foreground detection is facilitated by recording a few seconds of video with no subject at the beginning of each recording. The goal is to obtain a set of background images, which can then be used to create a model. The lighting conditions of the experiment—indoors with no incoming natural light—make it easy to create a working model, but more complicated scenarios would require a greater degree of care to build a model robust enough. The model chosen for the experiment is described by Zivkovic[17] as a series of improvements on the commonly used Gaussian Mixture Model; an implementation is readily available in the OpenCV library.

Recognition of the different leg segments is done most easily by colour. This is only possible if the recordings are carefully setup. The thigh, shank and foot should feature tight clothing in different colours that separate them from each other and the background of the scene. Dressing the subject in tight clothing that stands out from the background scene is a very common practice in marker-less motion capture, as it improves the segmentation from the foreground detection and avoids errors originating from the use of loose clothing on the subject. For this experiment, additionally differentiating each segment by colour eases the recognition step. An additional noise reduction step are performed to improve the quality of the results with the usage of simple filtering and morphology.

2.3 Synchronisation process

An important aspect to consider when obtaining movement data from different sources is synchronising them in time, frequency and space.

The synchronisation for the experiment is done manually and ad-hoc for the sake of simplicity, instead of implementing a more involved method that would allow higher speed and accuracy for bigger amounts of data. Such a method would ideally be added as hardware onto the employed systems, but a robust software solution is also possible. Sigal et al.[11] use a software implementation to synchronise the two systems employed in the HumanEva dataset.

Frequency sync is done based on the documentation for the hardware used, and was verified visually. The data source with the highest frequency was decimated to the frequency of the other source by an integer factor; therefore, no interpolation is performed.

The two systems are not synchronised in time; e.g.they start at slightly different times. Time sync is done visually by picking an easily identifiable moment in one system and finding it in the other one. To this end, some movements are performed in the recordings that would make this step easier and more precise, such as stomping quickly on the floor. This results in relatively unreliable synchronisation, but should be enough for testing goals, since the movements to be tracked are rather slow.

Space sync can be done in several ways, by establishing a common world origin using a landmark easily identifiable by both systems, or by determining one of the system's origin in the other system. In this case, it can be done by obtaining the position of the video camera according to the marker-based system, e.g. by placing one of the trackers in the position of the camera in a separate recording. Units must also be converted accordingly.

2.4 Joint angle calculation

Many different parameters could be extracted from the data depending on the interests of the application. For the experiment, the goal is to calculate the rotation of the two joints in the model: the knee and the ankle. This is a common goal for applications in biomechanics, where the global position of the body is not as relevant as the positions of some body parts in respect to each other.

The positions for the trackers have been chosen so that this task is very straightforward: because they are placed in pairs in a straight line for each leg segment, the vector between each pair is the direction of their segment. Figure 2.4 shows this simple process visually.

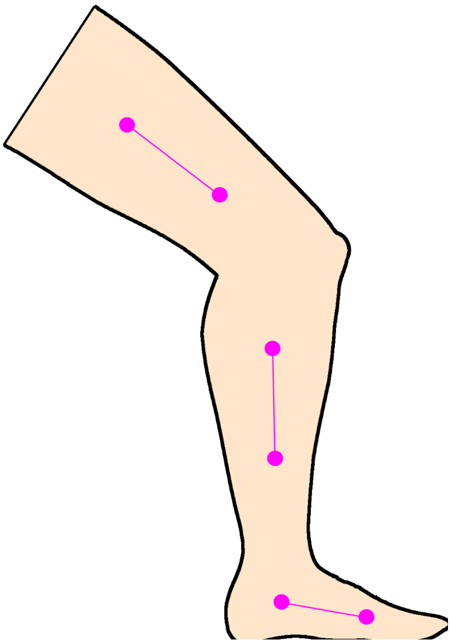

Figure 2.2: *The vector between the pair of markers of each segment is computed*

In the marker-less system, each of the three leg segments are fit into rectangles which then provide with a clearer orientation. Simple logic operations are done to ensure that the orientations are not flipped, e.g. a segment should be pointing forward and down instead of back and up. This is possible thanks to the restriction set on the experiment, but a more robust approach would be needed in an application with less predictable movements.

These more complete approaches use skeleton models to prevent unlikely poses, as has already been discussed.

The angle between two segments can then be easily computed from the dot product of their vectors u and v:

$$\theta = \arccos\left(\frac{u \cdot v}{\|u\|\|v\|}\right)$$

Figure 2.4 shows the computed angles between the segment vectors.

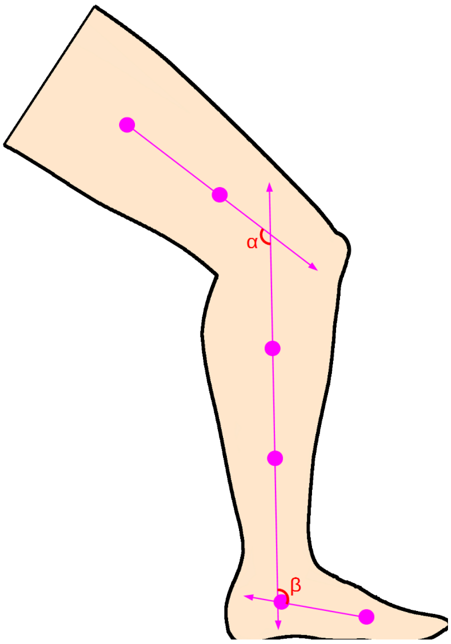

Figure 2.3: *The angles for the knee and the ankle are computed from the segment vectors*

2.5 Hybrid system

As explained previously, several methods exist for combining data acquired from different sensors. Since the two data sources chosen for the experiment are of similar nature, they can be combined directly for each frame after they have been aligned in space (through the synchronisation step).

This is a simplistic approach that ignores previous data and does not build a model to recover from missing data or noise. However, it would arguably present similar results for a small test to a more involved and generic approach, e.g. by using Kalman Filtering. In a practical scenario, a more robust tracking system would be greatly advised.

The data to be combined are the segment vectors which are extracted directly from the source. The angle computations would then be done using different combinations of vectors between both systems; one example of combination would be to combine the thigh vector from the marker-based system with the shank and foot from the marker-less.

2.6 Expected results

The recordings will be done with the full marker set, i.e. six trackers attached to the leg. As mentioned, different sub-sets of markers will be used for the different tests, ignoring the rest as if they were occluded or simply not present.

The expected results will be varying correlation levels between the marker-based and the marker-less systems depending on the leg segment; i.e.the marker-less system will behave better for specific leg segments. Therefore, different sub-sets of markers will provide different correlation levels when mixed with the marker-less data.

It is expected that these varying correlation levels coming from different subsets of markers will produce interesting results in terms of the trade-off accuracy and convenience that could be applicable for different application needs.

3 System description

In this section the different parts of the systems are described.The system is divided into three subsystems which process the marker-based and marker-less data and combine them in specific ways. The data flow and the interaction between the subsystems can be seen in figure 3.

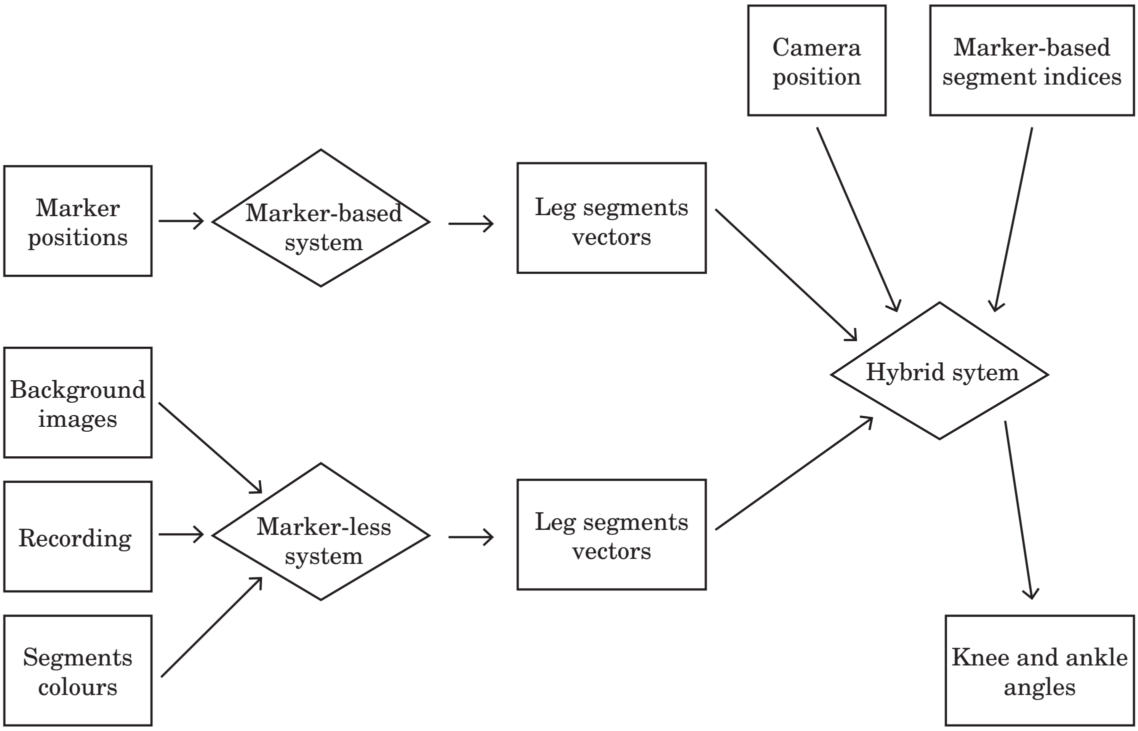

Figure 3.1: *Diagram of the different modules and data that compose the proposed system*

Marker-based It loads the marker data from the specified recording and computes the segment vectors.

Marker-less It trains the background subtractor with specified background frames, then segments each of the following frames. It performs colour segmentation to distinguish between the leg segments, aided by morphology operations and filtering to reduce noise. It then fits each segment's contour into a rectangle and computes the finds the direction of said rectangle, therefore obtaining the segment vector.

Hybrid It converts between the coordinate systems of the marker-based and marker-less systems using the annotated camera position. It then creates a new data array combining the specified segment vectors from the marker-based system and the marker-less system. Finally, it computes the angles for the knee and the ankle from these vectors.

4 Results

The proposed design has been tested on four acquired recordings. One colour camera and a set of trackers from an HTC Vive system were used, following the guidelines laid out throughout the document to ease the motion capture process; more specifically:

- The movement is performed roughly in a two-dimensional plane, walking on a straight line and moving the leg back and forth.
- The colour camera is placed so that it looks into the leg laterally, perpendicularly to the imaginary plane the leg is moving in. This way, it captures the whole movement.
- Before each recording, a few seconds are recorded without the subject in the image. This is used to train the background subtractor.
- All trackers must be visible by the tracking system throughout the whole section of the recording to be tested. The implementation tested does not attempt to track missing markers, although the marker-based system does recover from missing markers itself and re-identifies them properly.
- Tight clothing in different colours should be used to improve background subtraction and allow easy recognition of the leg segments.
- The position of the camera in the coordinate system of the marker-based system must be annotated.

The recordings feature different movements in two dimensions where the knee and the ankle angles exhibit a large number of combinations. The data were manually annotated to be resampled, cut and synced appropriately. In particular: the marker-based system records at 90 Hz and the marker-less system records at 30 Hz, and both start and end separately. The results from this manual process have been verified visually.

The number of frames containing only background is annotated for each video and the background subtractor is trained with the resulting frames.

Once these preliminary steps are completed, the recordings can be processed. The segments are identified and the pose is estimated, resulting in a three-dimensional vector for each segment in each frame, for each system.

The angle between the vectors associated with the thigh and the shank, and with the shank and the foot, are the final outcome of the motion capture process. These numbers are then compared between the systems.

The correlation indices between the two systems can be seen in table 4. Since both systems are processed in a simplified way with the most basic approach, the correlation indices are quite low.

The different combinations of thigh, shank and foot vectors between the two systems are then performed and the angles computed, resulting in the new correlation indices observed

Recording	**Marker-less correlation**		**Hybrid correlation with segment markers**					
			in thigh		**in shank**		**in foot**	
	Knee	**Ankle**	**Knee**	**Ankle**	**Knee**	**Ankle**	**Knee**	**Ankle**
1	**0.05**	**0.23**	0.07	N/A	0.10	0.00	N/A	0.11
2	**0.15**	**0.03**	0.34	N/A	0.12	0.16	N/A	0.07
3	**0.14**	**0.20**	0.29	N/A	0.09	0.09	N/A	0.27
4	**0.00**	**0.19**	0.01	N/A	0.06	0.44	N/A	0.28

Table 4.1: *Results from testing. Cells in green show improvement from original correlation; cells in red show worse correlation.*

in table 4. An improvement can be seen in most recordings for the combination of marker-based thigh with marker-less shank and foot, since the thigh was the segment with the most errors in the recognition step of the video.

5 Discussion

The results obtain agree with the experiment carried out by Becker et al.[8], even though this experiment did not make use of biomechanical models or more involved tracking techniques. Unfortunately, the over-simplification of the challenging marker-less segment recognition gave very poor correlation results in the first place, which invites to test further with a more robust marker-less approach.

Practical difficulties during recording raised unexpected problems to recognise leg segments without training data for an approach involving machine learning, and without a biomechanical model. In particular, there was trouble differentiating between the thigh and the shank due to problems with the video camera, which unfortunately affect both knee and ankle when the shank is misplaced. A more robust approach is advised, though time constraints did not allow for an appropriate implementation for this experiment.

5.0.1 Future work

As mentioned, a more robust marker-less approach should be used for further testing. This could be an implementation of a chosen method from the review by Yang et al.[6], such as the work by Corazza et al.[15].

More complete tests using a biomechanical model on a full body should be performed on a public dataset, such as HumanEva[11], to compare against other approaches. Although a challenging dataset, it is the most commonly used in marker-less motion capture research, as expressed by Yang et al.[6], and it should be possible to use it for testing hybrid systems in the same way as attempted in this work and by Becker et al.[8]: by using different combinations of marker subsets to constrain or correct the marker-less data.

A Appendix

A.1 Implementation details

The implementation to test the described design was carried out independently during this work.

It was implemented in the python programming language, making extensive use of the scipy and OpenCV libraries for numerical and computer vision methods, respectively.

The implementation was separated into the following modules:

parse_markers.py It loads marker positions from a local file or an FTP server. FTP was used during initial tests in order to avoid transferring each new batch of files between the recording system and the processing system.

plot_markers.py Creates an animation using pyplot with the marker positions. Used for visual testing.

leg_detection.py Uses the improved Mixture of Gaussians implementation in OpenCV (MOG2) for background subtraction and performs the marker-less segment recognition using colour segmentation, and morphology and filtering for noise reduction. It finally fits the contour of each segment into a rectangle and computes the segment vector from the angle of rotation of said rectangle. It also plays the recording, for visual testing.

camera_transformation.py Uses the annotated camera position to transform marker positions in global coordinates into camera coordinates within the Unity coordinate system.

calculate_lines.py Computes the joint angles from the segment vectors.

main.py Handles each module to go through all the recordings. It also attempts to reduce error by discarding frames with missing segments. Finally, it computes the correlation between the marker-less and the marker-based, as well as the different combinations tested in the hybrid system.

Extensive use was done of the ffmpeg toolset to annotate frames in the marker and video recordings, so they could be visually analysed and synchronised.

Bibliography

[1] D. Van Krevelen and R. Poelman, “A survey of augmented reality technologies, applications and limitations,” *International Journal of Virtual Reality*, vol. 9, no. 2, p. 1, 2010.

[2] Y. Baillot, L. Davis, and J. Rolland, “A survey of tracking technology for virtual environments,” *Fundamentals of wearable computers and augumented reality*, p. 67, 2001.

[3] H. Zhou and H. Hu, “Human motion tracking for rehabilitation—a survey,” *Biomedical Signal Processing and Control*, vol. 3, no. 1, pp. 1–18, 2008.

[4] T. B. Moeslund and E. Granum, “A survey of computer vision-based human motion capture,” *Computer vision and image understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[5] T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90 – 126, 2006, special Issue on Modeling People: Vision-based understanding of a person’s shape, appearance, movement and behaviour. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314206001263

[6] S. X. Yang, M. S. Christiansen, P. K. Larsen, T. Alkjær, T. B. Moeslund, E. B. Simonsen, and N. Lynnerup, “Markerless motion capture systems for tracking of persons in forensic biomechanics: an overview,” *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 2, no. 1, pp. 46–65, 2014.

[7] “Simi reality motion systems,” http://www.simi.com/en/, accessed: 2017-09-06.

[8] L. Becker and P. Russ, “Evaluation of joint angle accuracy using markerless silhouette based tracking and hybrid tracking against traditional marker tracking,” *Poster für Masterarbeit bei Simi Reality Motion Systems GmbH und der Otto-von-Guericke-Universität Magdeburg*, 2015.

[9] “Simi shape 3d,” http://www.simi.com/en/products/movement-analysis/markerless-motion-capture.html, accessed: 2017-09-06.

[10] “Simi motion 2d/3d,” http://www.simi.com/en/products/movement-analysis/simi-motion-2d3d.html, accessed: 2017-09-06.

[11] L. Sigal, A. O. Balan, and M. J. Black, “Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion,” *International journal of computer vision*, vol. 87, no. 1, pp. 4–27, 2010.

[12] M. Sandau, R. V. Heimbürger, C. Villa, K. E. Jensen, T. B. Moeslund, H. Aanæs, T. Alkjær, and E. B. Simonsen, "New equations to calculate 3d joint centres in the lower extremities," *Medical engineering & physics*, vol. 37, no. 10, pp. 948–955, 2015.

[13] J. Starck and A. Hilton, "Model-based human shape reconstruction from multiple views," *Computer Vision and Image Understanding*, vol. 111, no. 2, pp. 179 – 194, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314207001439

[14] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi, "Fusion4d: Real-time performance capture of challenging scenes," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 114:1–114:13, Jul. 2016. [Online]. Available: http://doi.acm.org/10.1145/2897824.2925969

[15] S. Corazza, E. Gambaretto, L. Mündermann, and T. P. Andriacchi, "Automatic generation of a subject-specific model for accurate markerless motion capture and biomechanical applications," *IEEE Transactions on biomedical engineering*, vol. 57, no. 4, pp. 806–812, 2010.

[16] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," *arXiv preprint arXiv:1402.5047*, 2014.

[17] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2. IEEE, 2004, pp. 28–31.