
Masterthesis Medientechnologie

RHOT-A real-time hand and object tracking system with low cost consumer grade hardware

vorgelegt von

Oliver Kalbfleisch

Erstgutachter: Prof. Dr. Arnulph Fuhrmann(TH Köln)

Zweitgutachter: Prof. Dr. Stefan Michael Grünvogel(TH Köln)

Juni 2018

Fakultät für
Informations-, Medien-
und Elektrotechnik

**Technology
Arts Sciences
TH Köln**

Masterarbeit

Titel: RHOT A real time hand and object tracking system with low cost consumer grade

Gutachter:

Prof. Dr. Arnulph Fuhrmann (TH Köln)

Prof. Dr. Stefan Michael Grünvogel (TH Köln)

Zusammenfassung: Was ein Abstract ist wird in der DIN Norm 1426 festgelegt: es ist ein Kurzreferat zur Inhaltsangabe. Die Definition des American National Standards Institute (ANSI) lautet: „An abstract is defined as an abbreviated accurate representation of the contents of a document“.

Es sollten 8-10 Zeilen Text folgen.

Stichwörter: Stichwort, Stichwort (hier sollten maximal 5 Stichwörter folgen)

Sperrvermerk: (optional) Die Einsicht in diese Arbeit ist bis zum TT. Monat JJJJ gesperrt.

Datum: TT. Monat JJJJ

Masters Thesis

Title: RHOT A real time hand and object tracking system with low cost consumer grade

Reviewers:

Prof. Dr. Arnulph Fuhrmann (TH Köln)

Prof. Dr. Stefan Michael Grünvogel (TH Köln)

Abstract: Hier sollte eine Übersetzung des obigen Abstracts auf Englisch erfolgen (und kein komplett neuer Text). Auch hier bitte die Begrenzung auf maximal 10 Zeilen Text ein-halten.

Keywords: (hier die Übersetzung der obigen Stichworte)

Restriction notice: (optional, falls Sperrvermerk oben ebenfalls existiert)

The access to this thesis is restricted until jdd month YYYY;.

Date: dd month YYYY

Contents

1	Introduction	1
2	Description of the hand in digital space	2
2.1	Physiological structure of the human hand	2
2.1.1	Constraints in hand motion	4
2.2	Kinematics	5
2.2.1	Forward Kinematics	6
2.2.2	Inverse kinematics	7
2.2.3	Jacobian-based methods	8
2.2.4	Singular Value Decomposition	10
2.2.5	Damped Least Squares	10
2.2.6	Newton Methods	11
2.2.7	FABRIK Algorithm	12
3	Tracking in real space	14
3.1	General tracking technologies	14
3.1.1	Optical tracking	14
3.1.2	Magnetic tracking	15
3.1.3	Acoustic tracking	16
3.2	Hand tracking Systems	16
3.2.1	Sensor based	17
3.2.2	Appearance Based - Optical marker	18
3.2.3	Model Based - Image analysis	20
3.3	State of the art technologies	22
4	Digital hand models	24
5	System prerequisites	26
5.1	Display	26
5.2	Tracking technology	27
5.3	Tracking Hardware	28
5.4	Data cleaning	29
5.5	Image analysis	29
5.6	Inverse kinematics	29
5.7	Hand model	30

6 System conception	31
6.1 Technical conception	31
6.2 Construction Details	32
6.3 Image analysis on the Raspberry	34
6.3.1 Image acquisition from camera	35
6.3.2 Image conversion to HSV colorspace	37
6.3.3 Mask construction and filtering	38
6.3.4 Contour finding and position calculation	41
6.3.5 object tracking	41
6.4 Systems communication	41
6.5 Data processing on the receiver side	43
6.5.1 Stereoscopic calculations	43
6.5.2 Depth calculation	45
6.5.3 Model position update	47
6.5.4 Inverse Kinematics	47
7 System Evaluation	48
7.1 Camera hardware and control	48
7.2 Image processing performance	49
7.2.1 Color accuracy and ROI size	53
7.2.2 Finger markers	54
7.2.3 Depth measurement accuracy	56
7.3 Inverse kinematics algorithm	58
8 Future Work	59
9 Resume	60
Abbildungsverzeichnis	ii
Tabellenverzeichnis	iii
Literaturverzeichnis	ix
10 Appendix	xii

1 Introduction

The standard interface between human and computer has for long years been mouse and keyboard. But with the advance of technology, new interfacing methods were developed in the last few years. Touch technology for interfacing with mobile devices and desktop computers has become a reliable technology and has been integrated into our everyday lives. Advances in capabilities of CPU as well as GPU hardware has built a foundation for the use of advanced augmented and virtual reality technology. 3D and stereoscopic rendering can now be accomplished even on mobile devices (with some limitations). For an intensely immersive experience, VR goggles are used to explore digitally created worlds. With this level of immersiveness, a touch device or just a mouse and keyboard setup is rather hindering for the user experience and the logical and more natural way of interfacing with our digital technology would be by utilizing the built-in human control elements that we carry around with us every day.

The human hands offer a form of control element that provides a large number of "*Degrees-of-Freedom*" (DOF's) and furthermore adds the possibility of combining these with a tactile component. The first more humble attempts of solving this problem already began in the early 1980's, where a lack of computational power and low-res imaging hardware made more complex systems impossible. These systems were mostly aimed at registering simple hand gestures like pointing as an interface method and were not able to reconstruct full hand positions in realtime[4]. Modern day elaborate system for hand tracking use infrared or stereoscopic Cameras to achieve a more performant tracking result, utilizing the calculation capabilities of standard consumer computers. These systems are mostly specialized for the task they are doing and are therefore rather expensive to attain.

This thesis assesses, if it is possible to create a real-time hand and object tracking system based on easily accessible consumer grade hardware and open source programs. The system should provide an experience for the user that is as natural as possible in terms of grabbing precision and the components should not hinder the motion capability of the hand. Furthermore the system should aim at being comfortable in terms of design as a cumbersome system would not be used in everyday life. Before evaluating a prototype system, an introduction to the anatomical properties of the human hand will be given. The understanding of these properties plays a major role in the setup for a hand simulation. As a follow-up, an overview of existing technologies in terms of tracking and display of the human hand will be given before the system conception, construction and analysis parts.

2 Description of the hand in digital space

Tracking of the human hand has always been a challenging problem. In comparison to other larger body parts like the arm or the head, the human hand itself contains a large variety of smaller parts, namely bones and muscles. These components have to be taken into account when trying to replicate the natural motion of the hand in digital space.

2.1 Physiological structure of the human hand

Lee and Kuni [30] described the human hand as "*an articulated structure with about 30 degrees of freedom [which] changes shape in various ways by its joint movements.*"

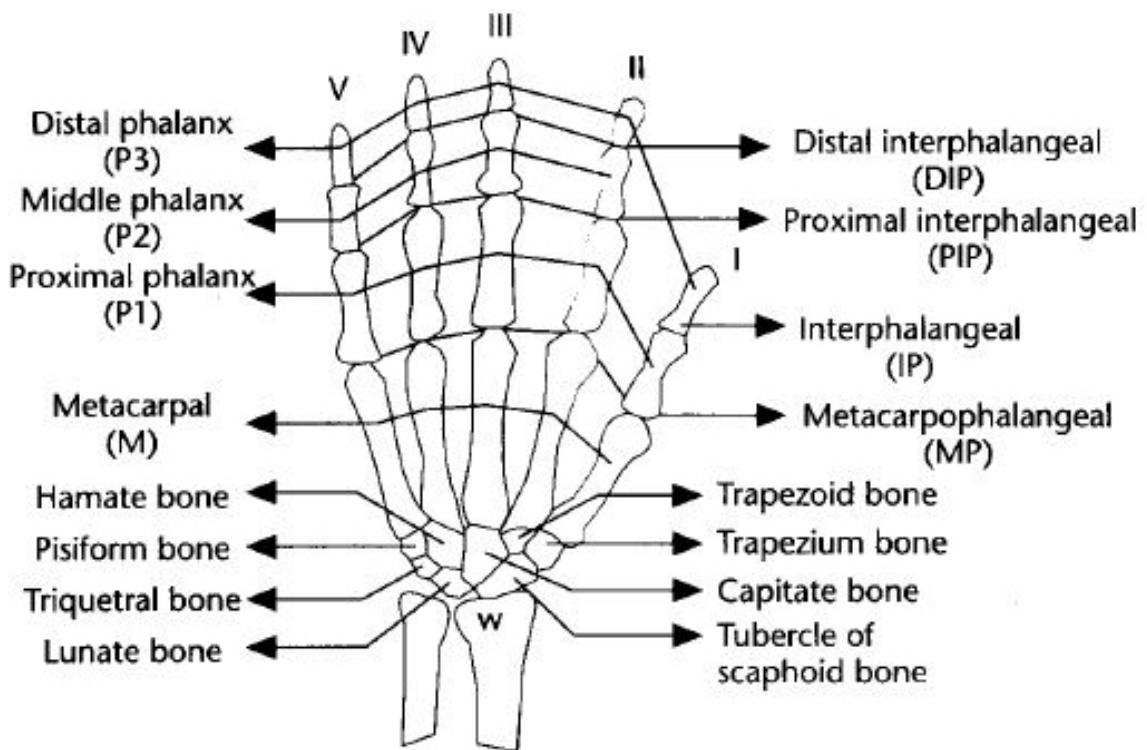


Figure 2.1: Bone structure of the human left hand ([30])

All of the hand parts are connected to at least one neighboring part via a *joint*. The *joints* affect the position of the connected part. To describe the movement of the hand parts, we can use the rotation angles of the joints to correlate to a specific position. To do so, we define a local coordinate system for each of the exiting hand joints. Through these coordinate Systems, we achieve a sequence of rotations in the local coordinate systems

of the joints. Such a sequence can then be used to describe a specific movement and/or position of a part. Not all of the joints in the human hand have equal *degrees of freedom*. Their functionality can be classified in the amount of DOFs (*degrees of freedom*)[25]

- 1 DOF
 - A joint movement that can perform a **flexion** or **twist** in one direction
- 2 DOF
 - A joint movement that can perform **flexion** in more than one direction (**directive**)
- 3 DOF
 - A joint movement that permits simultaneous **directive** and **twist** movements. (**spherical**)

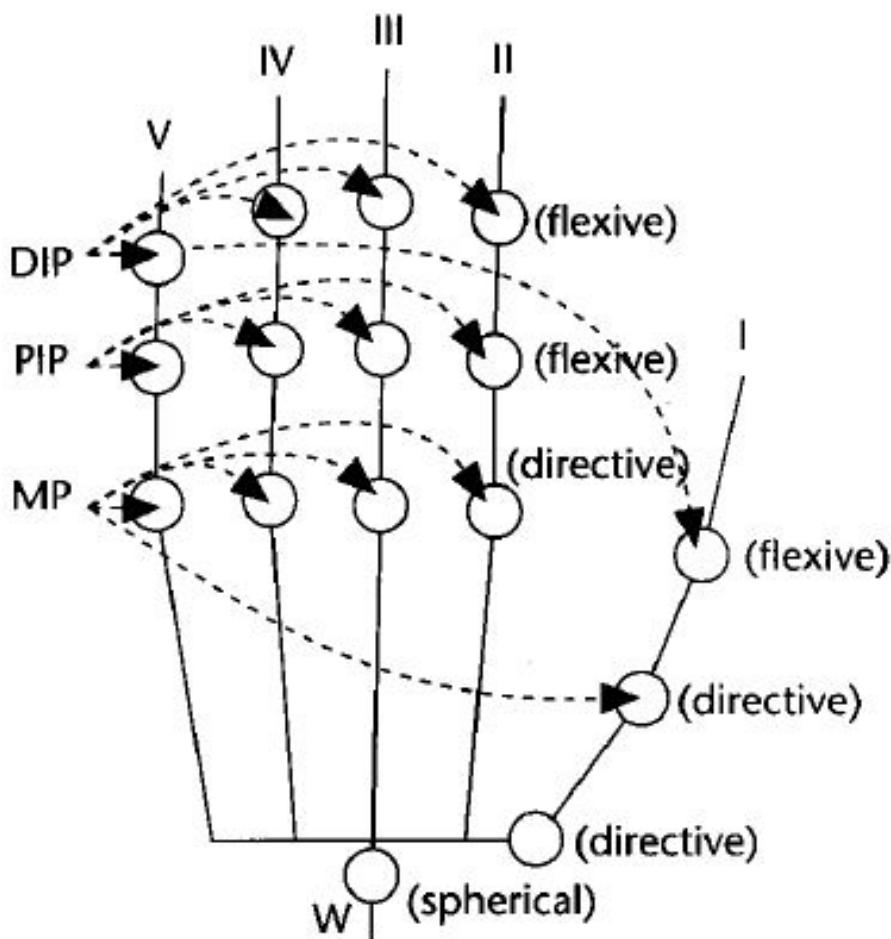


Figure 2.2: Types of degrees of freedom for the joints of the human hand

When looking at the DOF's displayed in Figure 2.2, each finger (II-V) sums up to 4 DOF's and the thumb to 5 DOF's. Also considering 6 DOFs for the rotation and position of the whole hand, the total sum add up to 27 DOFs.

2.1.1 Constraints in hand motion

A full usage of all the declared DOFs would lead to a large amount of possible combinations. Since the hand is not only made up of bones but also muscles and the skin, we can impose some constraints [3, 43] to the movement of the joints. Ling, Wu and Huang[31] proposed following classification for the constraints:

- **Type I constraints**

-A constraint that limits the range of finger motions based on hand anatomy

- **Type II constraints**

- A constraint that limits the position of the joints during finger movement

- **Type III constraints**

-A constraint that limits position based on natural hand motions

The **Type I** and **Type II** constraints rely on the physiological and mechanical properties of the human hand. **Type III** constraints are results of common and natural movements. Therefore these constraints will be differing from person to person. However, these movements are to a certain degree similar for everyone. Accordingly, a broad grouping can be applied. For example, the curling of the fingers at the same time when forming a fist is way more natural than curling each finger by itself. Here the motion of the hand is quite similar between different persons, but the constraints cannot be described in a mathematical form. The examples for the other two constraint types are less complicated. For the **Type I** constraint, there is no doubt that the position of the fingertip is limited by the length of the other finger segments and thereby can only reach as far as the combined length. If you want to touch your hand palm with your fingertip, all joints in the finger have to be bent to achieve this position, which is a constraint example for the **Type II**. For the mathematical description of the **Type I** and **Type II** constraints, the following inequalities can be used:

$$\begin{aligned} 0^\circ &\leq \Theta_{MP_{flex}} \leq 90^\circ \\ 0^\circ &\leq \Theta_{PIP_{flex}} \leq 110^\circ \\ 0^\circ &\leq \Theta_{DIP_{flex}} \leq 90^\circ \\ -15^\circ &\leq \Theta_{MP_{abduct/adduct}} \leq 15^\circ \end{aligned} \tag{2.1}$$

For the middle finger, a further specific constraint is applicable as the middle finger's MP joint normally does not abduct and adduct much. Therefore an approximation is applicable, resulting in the removal of 1 DOF from the model:

$$\Theta_{MP_{abduct/adduct}} = 0^\circ \tag{2.2}$$

The same behavior can be seen in the combination of hand parts labeled W(the connection point between hand and lower arm). This approximation also eliminates one DOF on the connected thumb:

$$\Theta_{W_{abduct/adduct}} = 0^\circ \quad (2.3)$$

Since the DIP,PIP and MP joints of our index, middle, ring, and little fingers only have 1 DOF for flexion, we can further assume that their motion is limited to movement in one plane.

The **Type II** constraints can be split into *inter-finger* and *intra-finger* constraints. Regarding *intra-finger* constraints between the joints of the same finger, human hand anatomy implies that to bend the DIP joints on the specific finger, the corresponding PIP joints of that finger must also be bent. The approximation for this relation[45] can be described as :

$$\Theta_{DIP} = \frac{2}{3}\Theta_{PIP} \quad (2.4)$$

Inter-finger constraints can be imposed between joints of adjacent fingers. *Inter-finger* constraints describe that the bending of an MP joint in the index finger forces the MP joint in the middle finger to bend as well.

When combining the constraints described in the above equations, the initial number 21 DOF's of the human hand can be reduced to 15. Inequalities for these cases, obtained through empiric studies, can be found in [30].

2.2 Kinematics

The preceding sections gave an overview of how we can describe a model of the human hand and introduced some limiting constraints. With the model and the constraints, we can now start to build a kinematic system for the animation of the model.

Kinematic systems contain so called *kinematic chains*, which consist of a *starting point* or *root*, kinematic elements like *joints*, *links* and an *endpoint*, also called *end effector*. Applied to the human hand, the whole hand model represents the kinematic system. This system contains several *kinematic chains*, namely the fingers of the hand with the fingertips being the *end effectors* of each of these chains.

Each of these components has its set of DOF's which can be described mathematically. The established convention for describing this kind of linked systems is the *Denavit-Hartenberg* convention (D-H)[9]. It was originally introduced in order to standardize the coordinate frames for spatial linkages in the field of robotics, but has since then been applied in other fields as well.[47]. As hand movements start, the states of the kinematic chains begin to change. Joint angles and end effector positions are modified until the end position is reached. To represent the new position and angle values of our physical hand with a kinematic system, two major paths for achieving a solution can be taken.

2.2.1 Forward Kinematics

The Forward kinematics(FK) approach uses the knowledge of the resulting angles and positions after the application of known transformations to the kinematic chain. The resulting new positions of the *joints* and *links* between the *root* and the *end effector* is used to solve the problem of finding the *end effector's* position.

We can denote the existing end effectors relative position to an origin as s_1, \dots, s_k . The s_i position is the result of the combination of all the joint angles in the corresponding kinematic chain. Respectively, we define the target position of the end effectors as t_1, \dots, t_i , with t_i being the target position for the end effector s_i . The required positional change for the end effector can now be described as $e_i = t_i - s_i$. In systems with more than one end effector, like our hand system, the components can be written as vectors.

$$\begin{aligned}\vec{s} &= (\mathbf{s}_1, \dots, \mathbf{s}_n)^T \\ \vec{t} &= (\mathbf{t}_1, \dots, \mathbf{t}_n)^T \\ \vec{e} &= \vec{t} - \vec{s}\end{aligned}\tag{2.5}$$

As the vector components of \vec{s} are results of the chain joint angles $\theta_1, \dots, \theta_n$ and therefore are effected by them, we define:

$$\begin{aligned}\vec{s}_i &= f_i(\boldsymbol{\theta}) \\ \vec{s} &= f(\boldsymbol{\theta})\end{aligned}\tag{2.6}$$

With $\boldsymbol{\theta}$ being the column vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$.The second vector equation displayed in (2.6) is also called the *Forward Kinematics*(FK) solution.

The advantage of an FK solution is that there is always an unique solution to the problem. In consequence, this approach is commonly used in the field of robotics, where the information on the chain elements is easily available. The tracking of the human hand and all of its chain components is rather complicated. Therefore a solution which takes a known position of the *end effector* and calculates the parameters for the rest of the chain is more desirable.

2.2.2 Inverse kinematics

"Inverse Kinematics (IK) is a method for computing the posture via estimating each individual degree of freedom in order to satisfy a given task" [1]

The concept of *Inverse Kinematics* (IK) already describes its principle in its name. It takes the reversed approach in comparison to the FK principle in section 2.2.1. Instead of knowing the states of the chain elements and calculating the resulting position of the *end effector*, we take the position of the *end effector* and try to retrieve the possible states of the other chain elements.

$$\boldsymbol{\theta} = f^{-1}(\vec{s}_d) \quad (2.7)$$

The result of this equation is the vector $\boldsymbol{\theta}$ for which the values of \vec{s} coincide to the desired configuration \vec{s}_d . In the case of an optimal result, this configuration would have the same position values as the target positions.

The main problem with this method occurs in the calculation of the f^{-1} function, due to it being a highly non linear operator which is not easily invertible. The approaches that [1] describes to counter this problem will be displayed later on in this chapter. In contrast to having a unique solution with the FK approach, the IK approach can end at the point of not finding a suitable solution. Figure 2.3 displays three possible outcomes for the IK approach.

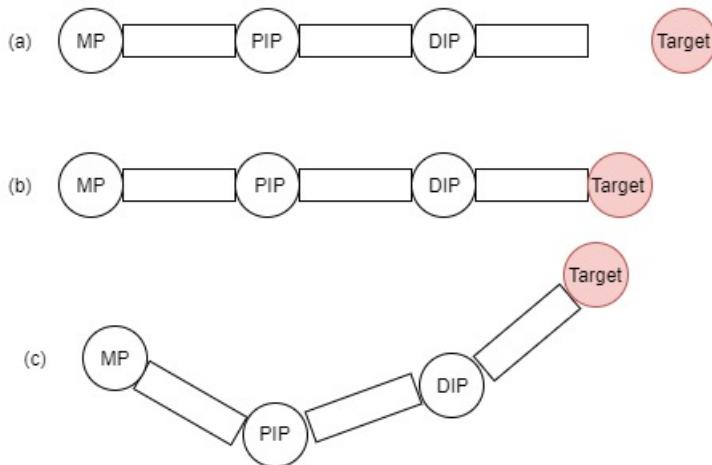


Figure 2.3: Possible solution for an IK problem of a human finger:(a)The given target position of the end effector can not be reached. (b) The given target can only be reached by one solution.(c) The target position can be reached with multiple different solutions.

2.2.3 Jacobian-based methods

Jacobian inverse

One common approach to solve the IK problem is the utilization of a Jacobian matrix and an iterative calculation process. This matrix contains the partial derivatives of the chain systems relative to the end effector \mathbf{s} . When using the Jacobian, a linear approximation of the IK problem will be applied for solving. The approximation's components model the end effector's relative movement to changes in transitions of the systems link translations and joint angles. Therefore, the resulting function is dependent on the joint angles $\boldsymbol{\theta}$ values and can be defined as:

$$J(\boldsymbol{\theta})_{ij} = \left(\frac{\partial \mathbf{s}_i}{\partial \theta_j} \right)_{ij} \quad (2.8)$$

$i=1,\dots,k$ and $j=1,\dots,n$. Further readings on methods for the calculation of Jacobian matrices can be found in [42]. Based on definition (2.8), an entry for the j -th rotational joint would be calculated as follows:

$$\frac{\partial \mathbf{s}_i}{\partial \theta_j} = \mathbf{v}_j \times (\mathbf{s}_i - \mathbf{p}_j) \quad (2.9)$$

Here \mathbf{p}_j is the position of the joint, and \mathbf{v}_j is the unit vector pointing along the current axis of rotation for the joint. Taking the derivative of definition (2.6) with respect to time gives the basic equation for forward kinematics that describes the velocities of the end effectors:

$$\dot{\mathbf{s}} = J(\boldsymbol{\theta})\dot{\boldsymbol{\theta}} \quad (2.10)$$

Now having all the values for the angles, the *end effector* position and the target positions, we can compute the resulting Jacobian matrix. Thereafter we seek an update value $\Delta\boldsymbol{\theta}$ for incrementing the current joint values:

$$\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{curr} + \Delta\boldsymbol{\theta} \quad (2.11)$$

The idea here is that the chosen value for $\Delta\boldsymbol{\theta}$ should lead to the resulting $\Delta\vec{s}$ being approximately equal to \vec{e} from (2.5). The \vec{e} can be approximated by:

$$\Delta\vec{s} \approx J(\boldsymbol{\theta}_{curr})\Delta\boldsymbol{\theta} \quad (2.12)$$

Using this approximation we can reformulate the FK problem as $\vec{e} = J\Delta\boldsymbol{\theta}$ and therefore our inverse kinematics problem form 2.7 can be expressed as $\Delta\boldsymbol{\theta} = J^{-1}\vec{e}$.

The problem we run into with this solution is the construction of the inverse Jacobian matrix. The Jacobian \mathbf{J} may not be square or invertible. In the case of it being invertible, the result may only work inferiorly because of it being nearly singular. Being singular means that no change in joint angle values may achieve the desired end effector position as an outcome.

Jacobian transpose

One approach to calculating the value of $\Delta\theta$ without having to calculate the inverse of \mathbf{J} is done by replacing the inverse with the transpose of \mathbf{J} .

$$\Delta\theta = \alpha \mathbf{J}^T \vec{e} \quad (2.13)$$

Of course the transpose and the inverse of \mathbf{J} are not the same thing. When using the theorems displayed in [7, 56] we can show that:

$$\langle \mathbf{J} \mathbf{J}^T \vec{e}, \vec{e} \rangle = \langle \mathbf{J}^T \vec{e}, \mathbf{J}^T \vec{e} \rangle = \|\mathbf{J}^T \vec{e}\| \geq 0 \quad (2.14)$$

Under a sufficiently small $\alpha > 0$ the updated angles from 2.11 will change the end effector positions by approximately $\alpha \mathbf{J} \mathbf{J}^T \vec{e}$. They also state that the value of α can be calculated by minimizing the new value of the error vector \vec{e} after each update.

$$\alpha = \frac{\langle \vec{e}, \mathbf{J} \mathbf{J}^T \vec{e} \rangle}{\langle \mathbf{J} \mathbf{J}^T \vec{e}, \mathbf{J} \mathbf{J}^T \vec{e} \rangle} \quad (2.15)$$

Jacobian pseudo inverse

Instead of calculating the normal inverse of the Jacobian, which can lead to the problems described before, we can use the so called *pseudo-inverse*[6] for the calculation. The *pseudo inverse* is defined for all matrices \mathbf{J} , even ones which are not square or not of full row rank.

$$\Delta\theta = \mathbf{J}^\dagger \vec{e} \quad (2.16)$$

The *pseudo inverse* represents the best possible solution for $\mathbf{J} \Delta\theta = \vec{e}$ in respect to least squares, but does suffer from instability near singularities. It has the property that the matrix $(I - \mathbf{J}^\dagger \mathbf{J})$ performs a projection onto the null space of \mathbf{J} . Therefore, for all vectors φ , $\mathbf{J}(I - \mathbf{J}^\dagger \mathbf{J})\varphi = 0$. Therefore $\Delta\theta$ can be set by

$$\Delta\theta = \mathbf{J}^\dagger \vec{e} + (I - \mathbf{J}^\dagger \mathbf{J})\varphi \quad (2.17)$$

for any vector φ and still obtain a value for $\Delta\theta$ which minimizes the value $J\Delta\theta - \vec{e}$. The pseudo inverse of \mathbf{J} can be derived as follows:

$$\begin{aligned}\mathbf{J}^\dagger &= (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \\ \Delta\theta &= (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \vec{e} \\ (\mathbf{J}^T \mathbf{J}) \Delta\theta &= \mathbf{J}^T \vec{e}\end{aligned}\tag{2.18}$$

Using $\vec{y} = \mathbf{J}^T \vec{e}$, we can now solve the equation $(\mathbf{J}^T \mathbf{J}) \Delta\theta = \vec{y}$

2.2.4 Singular Value Decomposition

The usage of a *singular value decomposition* is another method of calculating the *pseudo inverse* of a jacobian matrix. A singular value decomposition of \mathbf{J} consists of expressing \mathbf{J} in the form

$$\mathbf{J} = \mathbf{U} \mathbf{D} \mathbf{V}^T\tag{2.19}$$

For an $m \times n$ Jacobian matrix, \mathbf{U} is $m \times m$ and the columns are the *Eigenvectors* of $\mathbf{J} \mathbf{J}^T$. \mathbf{D} is $m \times n$, and the entries are zero except along the diagonal where $d_{ii} = \sigma_i$ with σ_i being the singular value and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0$. Singular values are the non-negative square roots of the *Eigenvalues* of $\mathbf{J} \mathbf{J}^T$ and $\mathbf{J}^T \mathbf{J}$. \mathbf{V} is $n \times n$ and the columns are the eigenvectors of $\mathbf{J}^T \mathbf{J}$. The construction of the pseudo inverse is done as follows:

$$\mathbf{J}^\dagger = \mathbf{V} \mathbf{D}^\dagger \mathbf{U}^T\tag{2.20}$$

The pseudo-inverse of the diagonal matrix \mathbf{D} is obtained by replacing each positive diagonal entry with its reciprocal [16].

2.2.5 Damped Least Squares

The *Damped Least Squares* method is applied for inverse kinematics as it avoids the singularity problems that can occur when using the pseudoinverse. It further provides a numerical stable method for selecting $\Delta\theta$. First occurrences for inverse kinematics can be found in [36, 52]. Instead of finding the minimum for $\Delta\theta$ as a best solution for the equation displayed in 2.5, the goal of this method is to find the value that minimizes the following quantity with $\lambda \in \mathbb{R}$ as a dampening factor.

$$\| J\Delta\theta - \vec{e} \|^2 + \lambda^2 \| \Delta\theta \|^2\tag{2.21}$$

which can be rewritten to the corresponding

$$(\mathbf{J}^T \mathbf{J} + \lambda^2 I) \Delta \boldsymbol{\theta} = \mathbf{J}^T \vec{\mathbf{e}} \quad (2.22)$$

and therefore

$$\Delta \boldsymbol{\theta} = (\mathbf{J}^T \mathbf{J} + \lambda^2 I)^{-1} \mathbf{J}^T \vec{\mathbf{e}} \quad (2.23)$$

The selection of a correct λ value is essential for this method, small values of λ give accurate solutions but low robustness to the occurrence of singular and near-singular configurations, while large values of λ result in low tracking accuracy even when a feasible and accurate solution would be possible[1, 5].

2.2.6 Newton Methods

The Newton family of methods is based on a second order Taylor series expansion of the object function $f(x)$:

$$f(x + \sigma) = f(x) + [\nabla f(x)]^T \sigma + \frac{1}{2} \sigma^T H_f(x) \sigma \quad (2.24)$$

where $H_f(x)$ is the Hessian matrix. The downside of this approach is that the calculation of the Hessian matrix bares high computational cost for each iteration as a result of its complexity. Like the approximation of the Jacobian inverse, several attempts to lower the computational cost have been made by utilizing an approximation of the Hessian matrix based on a function gradient value. Since the Newton methods are posed as a minimization problem, they return smooth motion without erratic discontinuities. The Newton methods also have the advantage that they do not suffer from singularity problems, such as that which occurs when finding the Jacobian Inverse.

2.2.7 FABRIK Algorithm

A more recent approach in the kinematics field is the *FABRIK* algorithm proposed by Aristidou and Lasenby[2]. As shown in section 2.2.2, most approaches all depend on computational intensive matrix operation like calculating an inverse and may have problems with matrix singularity.

The *FABRIK* algorithm does not depend on these matrix operations as it solves for the position of a point on a line to retrieve the new joint positions. This is done in an forward and also inverse solving approach, iterating these steps until the calculated position converges towards the target position from the tracking data. Figure 2.4 illustrates the steps that are contained in one iteration step. The chain joints are denoted as \mathbf{p}_i with the distance \mathbf{d}_i being $|\mathbf{p}_{i+1} - \mathbf{p}_i|$. The target point for the end effector is denoted as \mathbf{t} . Step (a) displays the starting point of the iteration. The joint positions are taken from either a previous iteration or from an initial calibration. But before calculations can begin, the algorithm has to check whether the intended target point \mathbf{t} is reachable for the end effector. This is done by measuring the distance between the root of the kinematic chain and the target point \mathbf{t} . This value is then compared with the sum of the distances \mathbf{d}_i .

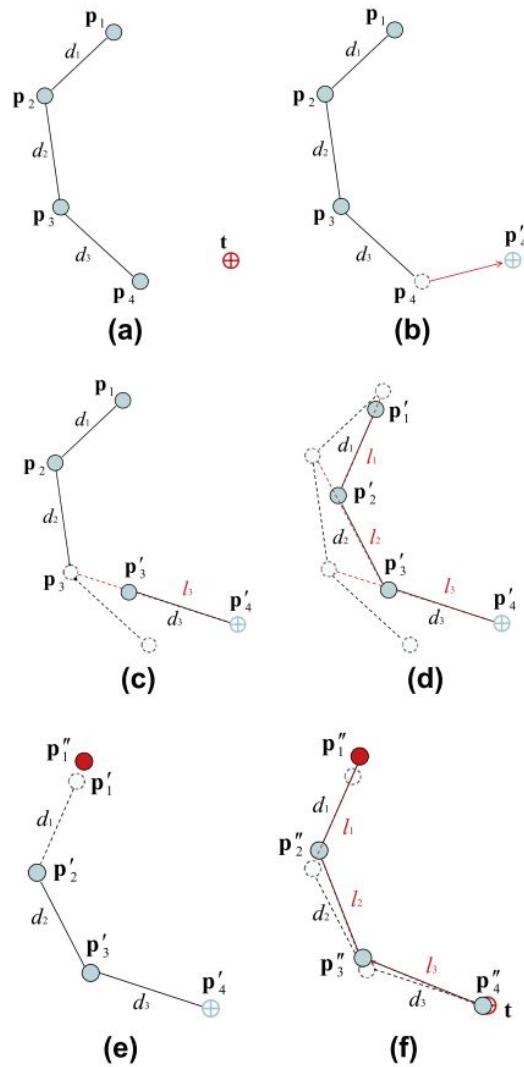


Figure 2.4: Forward and backward calculation steps for one iteration of the *FABRIK* algorithm.(a) initial position of the system.(b)End effector is moved to target position.(c) Determne position of next joint on constructed line.(d)repeat until root is reached. (e)move root joint to initial position. (f) repeat calculation in reverse direction [2]

$$dist_{t,d_1} < \sum_{k=1}^i d_i \quad (2.25)$$

If the summed distance is greater, then the target \mathbf{t} is within the reach of the system and the calculation can continue, otherwise the calculation has to be aborted and the error has to be managed otherwise. Assuming this requirement to be met, we can now begin with the first calculation. The inverse calculation step is the first step, which is displayed in **(b)** and **(c)**. The calculation is started at the end effector, moving to the root of the chain. Therefore we assume that the new position \mathbf{p}'_n with $n=4,\dots,1$ is equal to \mathbf{t} .

$$\mathbf{p}'_n = \mathbf{t} \quad (2.26)$$

From this new point, we can construct a line that goes through \mathbf{p}'_n and \mathbf{p}_{n-1} .

$$\begin{aligned} A &= \mathbf{p}'_n \\ B &= \mathbf{p}_{n-1} \\ \mathbf{l}_{n-1} &= \overline{AB} \end{aligned} \quad (2.27)$$

The resulting position of the new \mathbf{p}'_{n-1} point is located on this line with the distance of \mathbf{d}_{n-1} from \mathbf{p}'_n (see **(c)**).

$$\mathbf{p}'_{n-1} = \mathbf{p}'_n + \left(\frac{\overline{AB}}{|\overline{AB}|} \cdot \mathbf{d}_{n-1} \right) \quad (2.28)$$

Consecutively, this is done with the remaining joints until the root joint is reached (see **(d)**). This finishes the first half of the iteration step. With the calculated positions, we now perform a forward calculation, starting from the root until we reach the end effector. Since the root of the system normally does not move from its initial position, we have to reset the root joint to this value before starting to calculate the new positions of the subsequent joints (see **(e)**).

Analogous to the procedure in the inverse step, we construct the lines between the points and determine the new position values of the joints. The end result of this step is shown in **(f)**. At this point, we can decide if the result position of the end effector is appropriate in comparison to the value of \mathbf{t} . A simple threshold value for this case could be the position difference between these two points.

constraints
adding
for the
algo-
rithm

review
of algo-
rithm
capabil-
ties

3 Tracking in real space

The previous sections provided information on the structure of the human hand and how to represent it in the digital space. To apply the algorithms presented in Section 2.2 to our digital hand model, we need input data from our real world representative.

3.1 General tracking technologies

The methods for gaining positional data can be roughly categorized into two major groups, glove based methods and vision based methods. Glove based methods have already been in development since the 1980's [4] and have since then resulted in several solution attempts. Sturman and Zeltzer gave a survey on the existing tracking methods in their paper [49]. They distinguish between two areas of tracking, first the 3D positional tracking of the hand (and also other bodyparts) without regard to the hands shape and secondly the tracking of the hand shape with glove technologies. These tracking technologies presented are still applied today in modern tracking solutions [46, 55]. They account for solutions with optical tracking based on marker detection, magnetic detection via measurements of an artificial magnetic field [44] and acoustic measurements via triangulation of ultrasonic pings.

3.1.1 Optical tracking

The components for an optical tracking system are several cameras for object detection and some kind of tracking characteristic of the object to be tracked. These characteristics can be either artificially applied ones like active flashing infrared LED on key tracking positions of the body or infrared reflective markers. A series of cameras positioned around the tracking subject will then track these markers inside their visual fields.

The second method uses a single camera to capture the silhouette image of the subject, which is analyzed to determine positions of the various parts of the body and user gestures. The image data is supplied to special software which correlates the marker positions in the multiple viewpoints and uses the different lens perspectives to calculate a 3D coordinate for each marker. These image interpretation and correlation tasks require computationally costly operations. The marker tracking is also prone to errors through variation in lighting of the scene, material reflection properties and also marker occlusion as the trackers are moved. Also most of the systems rely on several tracking cameras for a complete coverage of the tracking space. This leads to a higher system complexity in terms of setup and calibration.

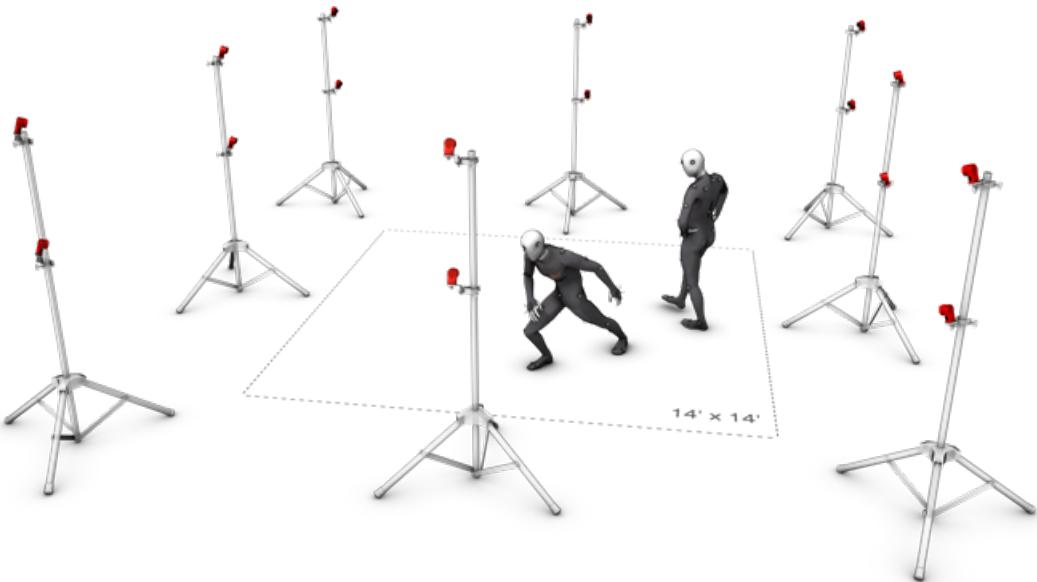


Figure 3.1: Example for an optical tracking system with passive infrared reflector markers [41]

3.1.2 Magnetic tracking

The usage of a magnetic field for position tracking is a relatively uncomplicated technique. The earth already provides us with a magnetic field to orient on. Similar to the optical trackers, magnet field sensors can be placed at key tracking positions. These sensors measure field strength in 3 orthogonal oriented axes to produce a 3D vector of the units orientation with respect to the excitation. As the earth's magnetic field is prone to changes based on geographical location, data corrections have to be applied to the measurements.

Another solution is to generate a local magnetic field via a multi-coil source unit. The source unit coils are energized in sequence and the corresponding magnetic field vector is measured in the sensor unit. With three such excitations, one can estimate the position and orientation of the sensor unit with respect to the source unit. The downside of this technology is that ferromagnetic and conductive material in the surrounding environment will have an effect on the generated magnetic field. These distortion effects will form small unwanted secondary source units through the induction of small *Foucault currents* from the main source's magnetic field.

Magnetic fields have the property of having an inverse cubic falloff as a function of distance from the source, which limits the range of operation for the system. Position resolution in the radial direction from source to sensor depends on the gradient of the magnetic field strength, and thus the positional jitter grows as the fourth power of the separation distance. In comparison to other tracking technologies, the magnetic field tracking solution has the convenience of not suffering from line of sight problems from tracker occlusion. The magnetic fields are capable of passing through the human body. Also the sensor size for measuring magnetic fields is rather small, giving the trackers a

small volume. Furthermore, only one source unit is needed for the tracking of multiple sensor units.

3.1.3 Acoustic tracking

The principles of acoustic tracking are very similar to those of the optic tracking technologies. Instead of using lightwaves, the systems utilize acoustic pulses of ultrasonic wavelengths to measure the time of flight between emitter and sensor for range measurement. To get a good measuring result from the systems, the used acoustic transducers have to be as omnidirectional as possible, so that the signal can be detected no matter how the emitter is positioned or oriented in the tracking volume. For the speakers to achieve a wide beam width, their size has to be small. To be able to build the microphones into the tracker, they can only have active surfaces a few millimeters in diameter. This leads to a reduction in range as the efficiency of acoustic transducers is proportional to the active surface area. Also acoustic systems can have problems with ambient noises occluding the signal.

This becomes even more critical when using such a system outdoors. Sound waves travel at a much slower speed than light waves which brings benefits and downsides with it. Sound waves can be reflected from objects, producing echos which arrive at the receiving sensor at a later point in time. Here the slower speed can be beneficial as the system can await the first sound occurrence to arrive at the sensor and filter out all later reflections from the data. The reflections of the previous pulse also have to be subsided before a new measurement can be made, lowering the update rates of the system. The air the sound waves travel through is also a limiting factor as humidity, air pressure and air currents can influence the traveling sound waves. In comparison to the optical systems, the acoustic systems are not as prone to occlusion errors since sound waves have a better ability to bend around obstacles than light waves. Most of these downside can be addressed with a combination of these systems with another form of tracking like in [13].

3.2 Hand tracking Systems

In comparison to the tracking of the whole human body, the tracking of the human hands with their maximum of 27 DoF's each in a small space is rather difficult to handle. The systems that try to achieve this have to encounter several difficulties. Self occlusion plays a great role in the system design, especially when working with only one tracking optical system for reduced complexity. Also these systems need to maintain a certain amount of processing speed for achieving a fluent reproduction of the captured motion. This demands the capability of processing large amounts of data in very short intervals. Most prototype testing for the described systems is done in a controlled environment where the background is known. For a more widespread use, these systems need to be capable of registering the hand on an unrestricted background, which may have a wide range

of patterns, color and lighting differences. Also human hand motion itself is quite fast, resulting in a higher frame rate demand for the tracking cameras. The following section will describe different approaches to solve the aforementioned difficulties.

3.2.1 Sensor based

Tracking systems that make use of sensors mounted to the hand via straps or a glove have already been in use since the 1970's. First prototype glove systems included the *Sayre Glove*[51] and the *Digital Entry Data Glove*[17].

The *Sayre Glove* is based off an idea of Richard Sayre. The system utilizes a light source and a photocell which are connected via a flexible tube. These components are mounted along each finger of the hand. The light that passes through the tube is influenced by the bending angle of the tube which corresponds to the finger bending. A large bending angle of the finger induces a larger bending angle of the tube, reducing the intensity of the light that reaches the photo diode. The resulting voltage in the photocell can then be mapped to a specific bending angle of the finger.

The *Digital Entry Data Glove* was patented by Gary Grimes in 1983. Instead of using only one type of sensors for measurement, this glove based systems used several sensor types for different measurements. Touch or proximity sensors were used for determining whether the user's thumb was touching another part of the hand or fingers. To measure the flexion of the joints in the thumb, index, and little finger, four *knuckle-bend* sensors were used. To get measurements for the tilting of the hand in respect to the horizontal plane two tilt sensors were mounted. Finally two inertial sensors for measuring the twisting of the forearm and the flexing of the wrist were utilized. This glove was intended for creating alphanumeric characters from hand positions. Recognition of hand signs was performed via a hard-wired circuitry, which mapped 80 unique combinations of sensor readings to a subset of the 96 printable ASCII characters.

These early systems only provided a limited amount of sensors and were rather unpractical in use. As they were developed to serve very specific applications, they were used only briefly, and never commercialized.

Newer developments in this sector include [19, 27, 32]. The *AcceleGlove* presented by [19] consists of six dual-axis accelerometers, mounted on the fingers and the back of the palm, reporting position with respect to the gravitational vector. Sensors are placed on the back of the middle phalanges, on the back of the distal phalange of the thumb, and on the back of the palm.



Figure 3.2: Acceleglove system with the described sensor positions [19]

Kuroda et al [27] introduced the *StrinGlov*, which obtains full degrees of freedom of human hand using 24 *Inductcoders* and 9 contact sensors, and encodes hand postures into posture codes on its own DSP. The bending angles of the fingers are measured with the *Inductcoders*.

These sensors relate the finger movement to a change in magnetic flux induced by the movement of the sensor parts that are attached to the finger. The sensor functionality is similar to the functionality of the light sensors from [51]. The 9 contact sensors, also based on magnetic fields, are put on the fingertips and on the inside of the hand to be able to measure contact between two fingertips or the fingertips and the hand. The system furthermore benefits from simple structure, which leads to low production cost.

Majeau et al [32] proposed a system that uses optical flexion sensors to determine the bending angles of the fingers. The optical flexion sensors consist of a LED, a photodetector and an optical fibre. The LED emits light which travels through the optical fibres and the intensity that results as the output from the fibre end is measured by the photosensor. This measuring principle also corresponds to the principle of [51].

Furthermore the system is also capable of measuring the abduction of two fingers. Here the LED is mounted to one finger and the detector on the other. The optical fiber is run from the LED down the finger to the knuckle base and then up to the detector. When abducting the two fingers, the angle at the knuckle base changes, resulting in an intensity change at the detector through the loss of intensity at the curvature of the optic fiber. The change in intensity of both measuring techniques can be mapped to corresponding hand movements. Further readings on sensor based hand tracking systems can be found in [10, 49]

3.2.2 Appearance Based - Optical marker

Optical approaches use properties of the human hand to estimate the current position. The evaluation of the hand pose is usually split up into two steps. The first one is the registration of the "real" hand.

This is mostly done by at least one camera which is aimed at the hand, supplying a continuous data stream to the evaluation software. In the second step, the evaluation software then tries to find key spots in the provided images. This data is then used for the search of the matching digital pose. The definition of these key spots can be achieved by several techniques.



Figure 3.3: Color glove setup used by Wang and Popovic [53]

An example is the usage of some kind of marking for finger segments as displayed in [11, 15, 53]. In the method described by Wang and Popovic [53], a color patterned glove is used. The glove is segmented into ten segments, colored randomly from a pool of ten distinct colors (see Figure 3.3).

The pattern is created by selecting twenty seed triangles on a 3-D hand model that are maximally distant from each other. The remaining triangles are assigned into patches based on their distance to each of these seeds. Each patch is assigned one of the ten colors at random. The jagged boundaries between the patches are artifacts from the low triangle count of the used hand model.

Algorithms that use other characteristics for tracking like texture or shading [28] rely on an accurate pose from a previous frame to constrain the posture search in the current frame. When using bare hand pose estimation two different hand poses can map to similar images. This can lead to an inaccurate pose estimation and the breakdown of the algorithm if the wrong estimation is made. The benefit of the color glove is that the unique patch pattern, hand poses always map to distinguishable images, simplifying the search process. Therefore it can effectively recover itself at each frame without the need of a previous frame. With the colored glove as a tracking feature, the images from the camera are prepared for a database sampling step. The database that was used consisted of 18,000 finger configurations.

A distance metric between two configurations was defined as the root mean square (RMS) error between the vertex positions of the corresponding skinned hand models. With this distance metric, a low-dispersion sampling was used to draw a uniform set of samples from the over complete collection of finger configurations.

The selected configurations are rendered at various 3-D orientations using a (synthetic) hand model at a fixed position from the (virtual) camera. The camera image was then compared to the database images. The comparison is done via a nearest neighbor lookup using a *Hausdorff-like* distance [21] and penalizing the distance to the closest pixel of the same color in the other image. The resulting pose from the database can then be applied to the digital hand model.

The method presented by Fredriksson and Ryen [15] also uses a color coded glove. In contrast to the fully colored glove described before, the used glove only has colored fingertips, color markers for the palm and a colored band for the wrist. The palm markers and the wrist band are used to retrieve the 3D position of the hand. The wrist marker is furthermore used to define the bounding box of the tracking area in a calibration process. Defining a bounding greatly simplifies further image analysis operations by taking away a large amount of unneeded image data.

After determining position and orientation of the hand in 3d by utilizing wrist band and palm marker position, the system tracks the grip angle and the lateral tilt angle for each finger. The grip angle is defined as the curling of the finger towards the palm. The second angle is defined as the lateral tilt angle which measures the spread of the fingers. The grip angle is calculated by comparing the density of finger pixels around an estimated knuckle line. From the density, an estimation for the fingers Y position around an origin position at the knuckles base is made. This value is then transformed into an angle value using an inverse tangent operation.

One downside of this tracking method is that it does not yet incorporate the movement of the thumb. The thumb has other movement possibilities and therefore needs a different algorithmic approach. Also the approach only focuses on the tracking of the hand data and does not yet implement a digital counterpart like the previously described system.

3.2.3 Model Based - Image analysis

Hand tracking with image analysis of optical marker positions has the flaw that these markers have to be put onto the hand in some kind of fashion. Using a glove for this purpose imposes just a little bit of inconvenience, but have the drawback of not fitting every hand shape equally and also having to deal with hygiene when switching between users.

The most natural way of tracking the human hand would be by simply using the properties of the human hand like skin color and shape for tracking. This would remove the need for any artificial added tracking features and have the greatest amount of comfort for the human user. A full overview of the technological advances in this area can be found in [34, 35]. The next part will only highlight some of the main topics displayed.

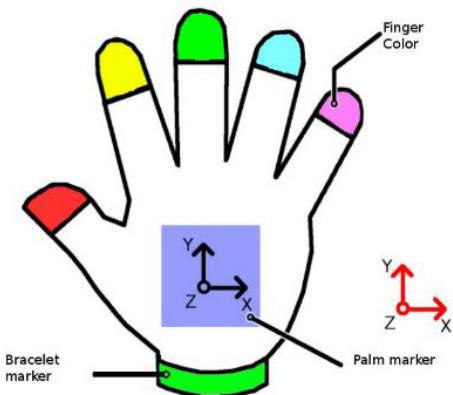


Figure 3.4: Color glove setup used by Fredriksson and Ryen [15]

As already explained before, the human hand motion leads to a large variety of shapes with many self-occlusions, which becomes even more critical when trying to reproduce the human hand pose via computer vision. Also the separation between background image information and the relevant hand information is a major topic, which is usually also the first step in the algorithmic process.

An established method for the extraction of background data was introduced by Stauffer and Grimson [48]. It uses the idea of representing each pixel value as a mixture of Gaussian functions (*MoG*) and updating these pixels during run time with new Gaussian functions.

The algorithm stores mean, variance, and likelihood value, based on previous values, for each pixel location and distribution. A new input is compared to the stored mean values for each distribution at this location.

A result that is less than a constant times the variance declares that this pixel is part of that distribution and it is updated accordingly. For a higher result, a new distribution is created which replaces the least probable distribution. A new input belongs to the background if the distribution that it is part of is one of the most likely distributions. A foreground Pixel that keeps the same color over time will slowly be incorporated into the background.

The decision rule, whether a new pixel belongs to an existing distribution or not, can be written as [26]:

$$if \ (|P_{new} - P_{mean}^d| \leq KP_{std}^d) \quad then \quad P_{new} \in Distribution \ d \quad (3.1)$$

where P_{new} is the new pixel value, P_{mean}^d and P_{std}^d are the stored mean value and standard deviation of distribution d and K is a constant. The convenience of this method is that it is also able to subtract background information from uncontrolled environments, making it possible for systems to work outside of lab conditions. After the segmentation of foreground and background is achieved, the hand pose has to be retrieved from the left over image data. Solutions for retrieving a full set of DOF as hand pose estimation data can be split into two main categories[12], *Model-based tracking* and *Single frame pose estimation*.

Model-based tracking refers to a top-down tracking method based on parametric models of the 3D hand shape and its kinematic structure. The root idea here is to execute a search at each frame of an image sequence to find the pose of the shape model that best matches the features extracted from the image(s). A prediction based on previous motion and dynamics data from the object is set up as a base for the search. Optimization features can be incorporated, like employing a local search around the calculated prediction to produce a single best estimate at each frame. The downside of this method

represents itself in the weakness of accumulating errors over longer run time periods due to occlusions and complexity of hand motion. These can be overcome by keeping track of multiple outcome hypotheses at each frame for error reduction.

Single frame pose estimation in contrast to the previous method, does not utilize assumptions on time coherence. At each iteration of the search algorithm the whole data set is searched for a match. This on one side introduces more computational work for achieving a match, on the other side lowers the amount of constraints that have to be put on the user when initializing the tracker system. Furthermore, the possibly rapid motion of the human hands and fingers acts as factor failure for time coherence dependent approaches, where a single frame approach is not as vulnerable.

3.3 State of the art technologies

The currently most used type of system for tracking hardware is a setup of a combination of RGB and infrared sensors for imaging and depth measurement (RGB-D). The first commercial system using this technology is the *Microsoft Kinect* system[58]. The system uses a infrared projector and an infrared sensor to measure distances in the captured scene.

Newer systems that also incorporate depth

sensing through infrared distance measurement are the *Intel RealSense* system and the more consumer focused leap motion hand tracker. The *Leap Motion Controller* does not supply a separate RGB sensor and does therefore only provide depth measurement data within a specific range from the device[54]. It combines two infrared cameras with three infrared emitter LED's, placing the system into the category of stereoscopic tracking systems.

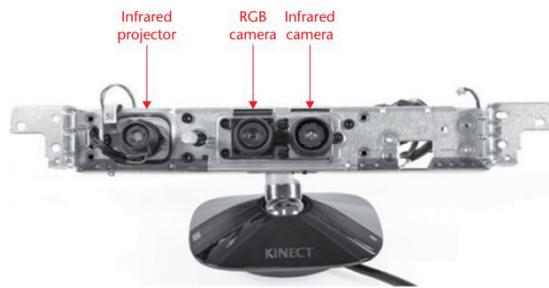
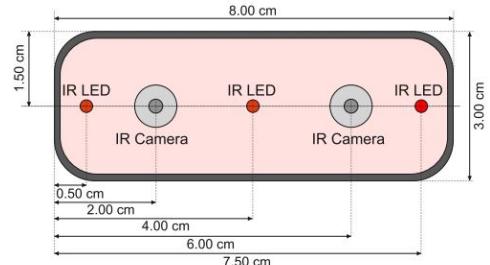


Figure 3.5: Microsoft kinect camera system with the three used sensor units[58]



(a)



(b)

Figure 3.6: Visualization of a (a) Real (using Infrared Imaging) and (b) Schematic View of Leap Motion Controller[54].

3 Tracking in real space

The latest product in this category are the *Intel Realsense* camera systems[22]. The Intel systems, like the *Leap Motion Controller* utilize a stereoscopic infrared system setup, and can also provide RGB Sensor data, depending on the selected model.

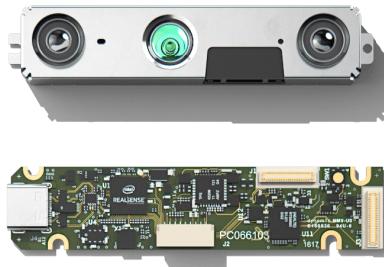


Figure 3.7: Visualization of a *Intel Realsense* depth module[22]

The *Kinect* can supply depth data at around 30 fps, it's modern descendants from Intel can supply data at frame rates of up to 90 fps. The systems provide the convenience of not having to apply any form of distinguishable marking to the surface that has to be tracked. The downside of this feature is that the devices only supply SDKs for accessing the collected data from the device. The post processing of this data needs more sophisticated image analysis or neural network processing to figure out feature points of the tracked object from the supplied image data[23, 38].

Image analysis of large image can be costly for real time applications and the training of neural networks for the analysis can consume a lot of time and tweaking until the trained result fits the application.

An recent approach of combining RGB-D Data and the usage of CNNs (Convolutional Neural Networks) to estimate hand pose and render a correct digital representaion of the hand is shown in [14].

Their approach uses two subsequently applied CNNs to localize the hand and regress 3D joint locations. The first CNN estimates the 2D position of the hand center in the input. The combined information of the hand position together with the corresponding input depth value, is used to generate a normalized cropped image. this image is then fed into a second CNN to regress relative 3D hand joint locations in real time. Pose estimation is furthermore refined by using a kinematic pose tracking energy to achieve more accuracy, robustness and temporal stability.

They also introduce a new photorealistic dataset that uses a merged reality approach to capture and synthesize large amounts of annotated data of natural hand interaction in cluttered scenes for CNN training data.

4 Digital hand models

After retrieving all needed data from the real world counterpart and calculation for pose estimation is finished, the returned solution has to be displayed in the digital world. Therefore a digital hand model is needed to represent the calculated data. Modern day computer animation techniques make it possible to create nearly photo realistic digital replicas of human skin wrapped onto anatomically precisely modeled body parts. These highly complex models acquire a lot of calculation resources and are mostly not appropriate for a mealtime rendering approach.

When adapting calculation data to digital hand models, those of lesser complexity are preferred, sacrificing the quality of the resulting output for speed and more fluent motion results.

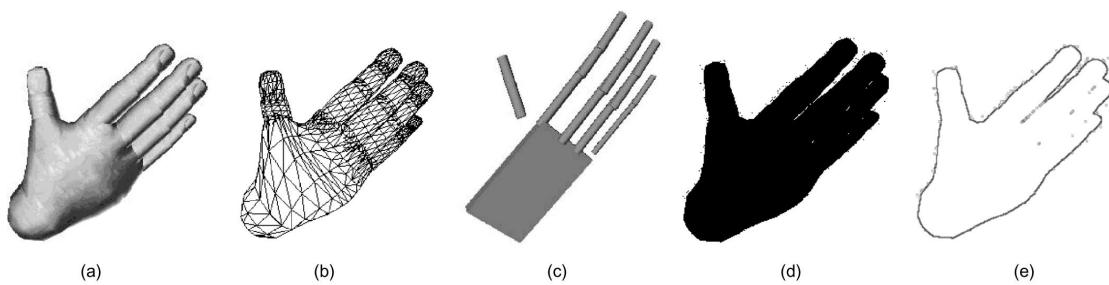


Figure 4.1: Pavlovic, Sharma et al: Hand models. Different hand models can be used to represent the same hand posture. (a) 3D Textured volumetric model. (b) 3D wire frame volumetric model. (c) 3D skeletal model. (d) Binary silhouette. (e) Contour. [43, p. 682]

Figure 4.1 shows different model representations of the same hand posture. The models of **(a)** and **(b)** represent the aforementioned types of models that would be used in modern computer-graphics for displaying hand models realistically.

These models are based on complex 3D surfaces like *NURBS* (nonuniform rational B-splines) which are wrapped around an underlying skeleton. The wire frame model of **(b)** give a good overview of how many vertex calculation have to be made for the surface in each frame, causing these models to hardly run in real time if not supplying a large amount of computational power.

In applications where the 3D orientation of each individual finger is not necessary for processing and the general silhouette is sufficient (e.g. general hand position tracking or sign language detection), even more reduced representation forms like **(d)** and **(e)** can be chosen. The model displayed in **(c)** is a simplification of the human hand structure with geometric primitives like cylinders and boxes. These geometric primitives have the

4 Digital hand models

virtue of being much easier to describe than a *NURBS* model. A cylinder for example is fully described by its height and its radius under the assumption that the position and orientation data is generally needed for all the described models. This type of model is also referred to as *skeletal models* as they mostly try to reproduce the structure of the human hand as described in Section 2.1

5 System prerequisites

Before elaborating a final system conception, some prerequisites have to be defined.

System components to be defined:

- Display hardware
- Tracking Technology
- Tracking Hardware
- Data cleaning
- Image analysis
- Data cleaning
- Inverse kinematics algorithm
- Hand model

5.1 Display

The first definition that has to be done is actually the last component of the processing chain. The final display hardware properties have to be defined initially as these are the determining factor for all other components. Since the goal of hand and object tracking is the most immersive when using a virtual reality headset rather than a normal display, these do not need to be taken into consideration.

Another device category which may be in the focus of interest are current mobile device platforms like *Google cardboard* or the *Samsung Gear* platform. In terms of creating a low cost consumer grade solution, these devices should be the first category to look at, as they are more widely available and do not need extra hardware except the smart-phone. Most modern devices have native support of WebGL in the browser, making it fairly easy to display VR content. They also come with the convenience of already having a set of integrated sensors for position and rotation integrated. The downside of this device class reveals when taking a look at the hardware specifications of these devices. Although the newest high tier consumer grade devices do have a good performance output for their size they are still no comparison for a desktop setup. Also these devices, as they are first of all mobile phones, show lacks for wire connections, making it necessary to mostly communicate over wireless protocols. This leads to a growth in infrastructure(server,

wi-fi-hotspot, etc.) and every further node in the processing chain leads to a physical signal delay.

After these first preliminary steps, only pure VR-Headsets like the *HTC Vive*[20] or *Occulus Rift* [37] should be the main focus. These headsets also provide the capabilities of motion tracking while utilizing the processing capabilities of a dedicated GPU and high-power CPU capabilities of modern desktop computers. Furthermore, they provide a variety of connection ports for wired connections, eliminating the need for wireless transmissions. In terms of cost, these devices do lie on the more costly side as you need the headset itself and a hardware setup which is capable of providing the output frame rate needed. But with the latest generations of GPU's and CPU's, this category was opened towards the normal consumer.

The *HTC Vive* as well as the *Occulus Rift* are able to display 90 images per second on their displays with each having a resolution of 2160x1200 pixels, resulting in a little over full HD resolution per eye. The duration of one frame is therefore around 11 ms. This is the time frame in which the rest of the components can theoretically supply the image data, do the image analysis to retrieve the tracking data, recalculate the IK model, calculate the object position and render the whole scene before sending it to the headset to be displayed.

5.2 Tracking technology

Physical sensors placed on the hand tend to give more precise tracking results compared to optical methods. The cost for the data precision is that the sensors have to be mounted to the hand in some kind of fashion. Cables or fibers, depending on sensor type, can hinder the hand movement. Utilizing a glove based system brings the downside of the "one size fits all" problem. Human hands can differ largely in size, therefore a glove will not be able to be used for a wide range of users. Furthermore cloth gloves tend to get dirty while usage and are therefore unsuitable in terms of hygiene.

Color markers can be utilized very easily. The simplest form could be colored electrical tape wrapped around the fingertips. These markers can be made disposable after usage. The size of the marker can be adapted to be large enough to realize a robust tracking while being small enough to not constrain hand movement. They also provide the possibility to perform a person specific calibration of the hand model in an initialization step. The easiest way to detect color markers is to use an optical tracking system. Monocular optical tracking systems are relatively easy to set up as they only need one camera, the complexity in the later processing steps of evaluating the marker positions from only a single frame rises drastically.

Stereoscopic systems in comparison have a higher initial effort in setting up and calibrating the system, but the further processing steps are fairly easy.

5.3 Tracking Hardware

As explained in the previous chapters, the motion of the human hand can be complex and in some cases really fast. Therefore, the optical tracking hardware should ideally be able to record images with the same or higher frequency as the display medium that is to be used. Prosumer and professional grade cameras are theoretically able to produce these kind of Frame rates (60 fps and above), but most of the time store these high frame rate videos directly to a hard drive rather than broadcasting them somewhere as most of the current day devices are not able to display such high frame rates yet. Also the hardware that is needed to record such specific high frame rates is relatively expensive. Consumer grade cameras like the GoPro that are affordable, can record at a Frame rate of 120 fps but the live transmission of this material has a time lack of several frames.

When trying to utilize low cost hardware the limitation of recording and sending a video signal from a camera to a processing hardware will therefore be limited to 60fps. This process would also introduce further latency on the transport and would force a further reduction of frame rate to satisfy the time frame given by the display frame rate.

Another option that can be taken into consideration is to completely eliminate the need to transport the image data from camera to another device. This can be achieved when the camera and the computational hardware are located on the same hardware.

The *Raspberry Pi 3* microcomputer, which comes from the internet of things world provides this capability. It combines a for its size quite powerful computation capacity with an on-board hardware connector for a camera. The camera hardware is able to record in a FullHD resolution at 60fps and at lower resolutions up to 90fps. The camera module is directly connected on the the chipboard, providing direct access for further processing. For a complete 3 dimension space tracking, a one camera setup is not sufficient as it lack the data for the depth position of the object. To get this data, further hardware is needed. One option would be the utilization of a time of flight based depth sensing system, but these system do come at a rather high entry level price.

Another option could be the tracking pods that HTC offers as an expansion to it's vive system. These trackers could provide 3D positional data but with the downside of being relatively bulky. This would hinder the movement space of the hand and the tracker would have to be positioned on the arm as near as possible to the wrist. This would result in the position of only the wrist and not providing data on each individual finger of the hand. But as hindering as these trackers are on the hand, they could be easily used to provide tracking data for used objects. the tracker could be positioned onto the object where it does not hinder the usage.

Since the hardware setup of the Raspberry Pi with the camera is relatively easy and cost efficient, the usage of a second Raspberry Pi with a camera is good leverage point. The two cameras can be setup to create a stereoscopic camera setup which provides informa-

tion about depth position through the disparity of the two camera images. Furthermore the image operations that need to be independently applied to each of the cameras frames can be done on the two separate device, eliminating the need for sending data traffic with image information through the network for post processing.

Downside of this method is that with all stereoscopic camera setups, the two cameras need to have some kind of frame synchronization to get the right two frames for evaluation.

The setup is not only meant for tracking the hand, but also for tracking objects in the tracking space, the system needs to be capable to do this task as well. Since the computational power of the raspberry is limited at some point, it is to be evaluated if a separate hardware component is needed for the object tracking or if the tracking can be integrated into the tracking procedure of the hand markers.

5.4 Data cleaning

Every system that deals with the computation of live data is suspect to data fluctuation in input and output values. It is also to be assumed that the system will not be able to achieve an ideal synchronization between the cameras. This will probably introduce a degradation of tracking performance which has to be handled. It has to be evaluated at which point a data cleaning will lead to the best end results.

5.5 Image analysis

The images recorded by the camera system need to be processed further down the line to get the positional data of selected hand features. The most common library to do such kinds of image processing is the *OpenCV*[40] library which is originally written in C and was ported into several other programming languages. It has build in features for camera calibration, color and object detection as well as several image optimization features that might be helpful for the processing.

5.6 Inverse kinematics

The inverse kinematics model is the second essential part of the setup after the tracking hardware. It receives the position data from the image post-processing and calculates the resulting current hand model position. The applied algorithm for the IK solution should be able to keep up with the data-flow from the tracking algorithm and should produce a fluid optical result on the display. As the calculation of the IK solution and the rendering of the scene with the models should be done on the same machine, computation power and memory consumption should be critical criteria. As the system will rely on optical

tracking data, an algorithm that is capable of dealing with tracker occlusions and recovery from this data loss should be prioritized.

5.7 Hand model

For evaluation purposes, a rather simple model form should be used for the representation of the tracked hand in the digital space. The outcome results for tracking accuracy, update speed and the resulting match of physical and digital position of the hand are of more relevance than a highly realistic hand model.

To receive a correct optical result of the hand model finger and hand dimensions have to be measured for each hand specifically and stored into a data model for correct bone dimensions of the hand model. Ideally, this step should be automateable and included into the system spin up process.

6 System conception

6.1 Technical conception

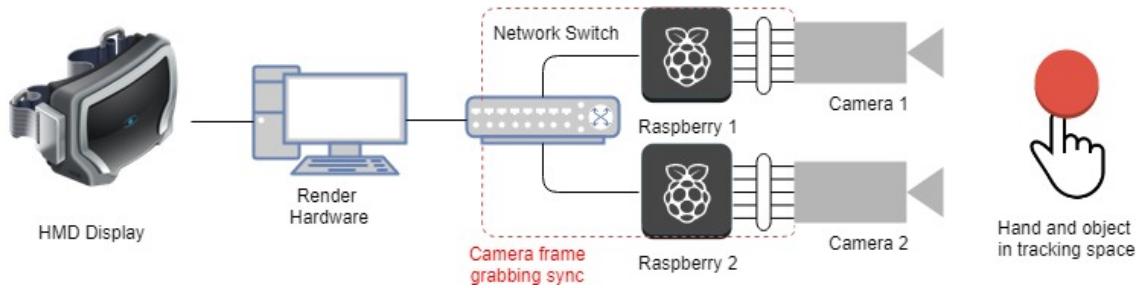


Figure 6.1: Technical conception of the single hardware parts

As depicted in Figure 6.1, the technical components of the setup are relatively simple. Instead of using one large dimensioned unit which takes care of all calculations, the image processing steps that have to be done on both stereo images anyway, are outsourced onto the Raspberry Pi's. The image processing can be done on the two Pi's in parallel, which cuts down processing time. Furthermore the image data stay on the device and does not have to be sent to a min unit which would introduce network and data reading latency.

edit
figure
above

The two **Raspberry Pi 3 Model B**, which are used as the controllers for the **Raspberry Pi Camera** are connected via network cable to a network switch. A connection over wireless network would be possible with the on-board capabilities, but this might create latency problems and/or signal interference with other existing networks. Therefore a cable connection is the safer solution. The switch also connects to the "Master" PC to which the "Slave" Pi's communicate their data. The Master also takes care of the stereoscopic calculations, as it is dimensioned with far more computational power than the slaves. The 2D positional data from the slaves and the calculation results from the stereo image disparity is fed into the hand model running on the "Master". The model solution is then applied to the digital hand model and rendered.

6.2 Construction Details

For the physical construction of the system, it is important to know what the primary use case of the system will be. A mobile system is principally not impossible as the Raspberry Pi's could be run with battery power. A wireless communication leads to the beforehand described Problems and a long enough cable connection is rather cumbersome. Furthermore the cameras need to be calibrated at a fixed position for correct depth measurements. Some kind of mounting harness would be needed where the cameras stay in a fixed position.

A much easier approach for the first prototype is defining a seated usecase for the user. Since the systems main chore is to track the hand of the user, a seated position reduces position tracking errors as one does not have to take account large positional changes of the rest of the body. The tracking volume that needs to be accomplished for a seated position is also reduced by our physical limitations of the arm length. A tracking volume of about $100 \times 100 \times 100$ cm should be sufficient to track most of the movement.

A frame of 30x30 mm construction Profile is used as base to mount all the components of the system. To reduce possible color reflections from the background of the tracking space, black fabric is used as background material. The other components of the system are:

- 2 Raspberry Pi 3 Model 3 with 1GB RAM and a 64GB microSD Card running a current Raspbian OS and a C++14 compiler
- Compiled version of OpenCV and Boost librarys
- Java LWJGL 3 and CALIKO library
- 2 Raspberry Pi Camera Modules V2 with a 100 cm connector cable
- Ethernet Switch
- Monitor for calibration and debugging of the Raspberry side software
- PC capable of supporting a HMD and a current Java version 8+
- Mounting peripherals
- Active cooled case for the Raspberry to help with heat reduction
- Two micro USB power supplies with 3.0 A output current

The active cooled Raspberry pi cases are an optional part of the construction, but the real time image processing on the CPU of the Raspberry together with the running camera produces larger amounts of heat than in normal operation mode. The applied heat sinks and the active cooling fan help with faster heat dissipation. The 3.0A power supply units

are recommended as lower output currents still work but can lead to a power brown out on the system. This can cause system damage over longer usage times and even make the used SD card unreadable.

Mounting peripherals

The used cameras are basically the camera sensors mounted on a pcb with an optic. They have 4 drilling holes as fixations points. A simple holder plate was constructed digitally and printed with a 3D Printer for faster prototyping. The holder was designed as two parts to be able to change the mounting distance of the camera setup from the construction profile if needed. The two parts are held together by 4 screws for permanent fixation. Cut outs in the mounting plate for the cameras provide the ability to mount the cameras with two machine screws and move to end position of the cameras along a fixed axis (see Figure 6.2). This ensures that the cameras are mounted on one axis. It furthermore give the opportunity to change the interaxial distance of the cameras if needed. for the current setup this distance will be set to 75 mm which corresponds to the interaxial distance of the human eyes.

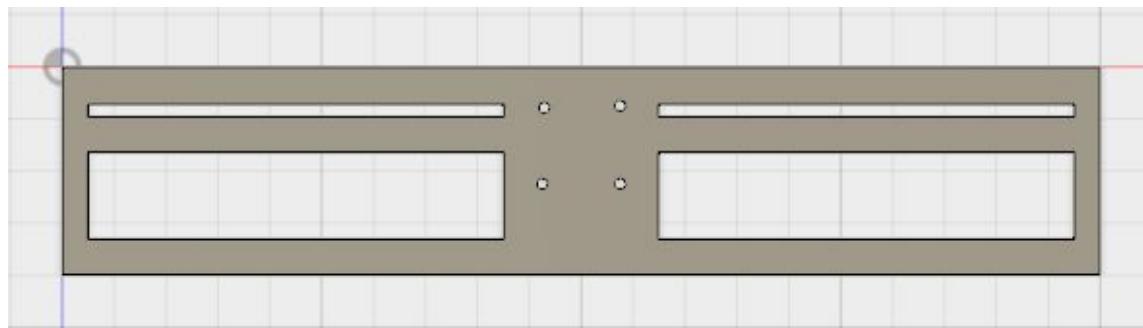


Figure 6.2: Technical drawing of the camera holder

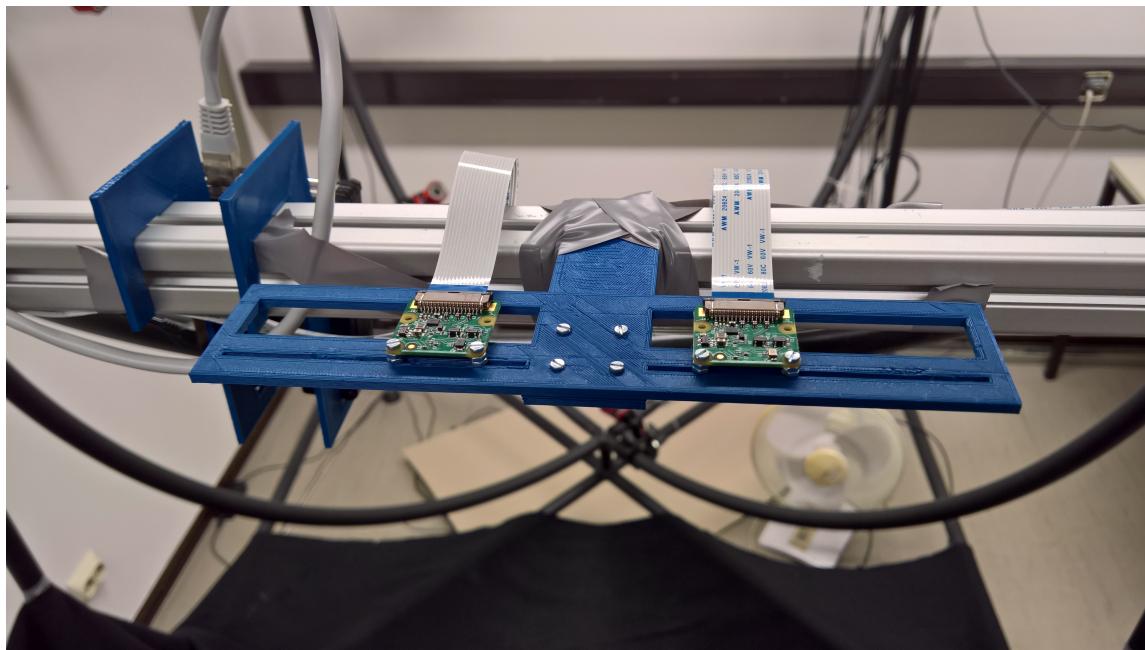


Figure 6.3: Printed camera holder with mounted cameras

6.3 Image analysis on the Raspberry

For the image analysis part, *OpenCV3* with it's *Python 3* bindings is chosen. The *OpenCV* package needs to be downloaded and compiled onto each device before it can be used. Before the actual image analysis part can take place, the used cameras need to be calibrated. As they are basically pinhole cameras, they introduce amounts of radial and tangential distortion to the images. These distortions need to be compensated for, especially when using them as source for the stereo images. Image distortion in these pictures would lead to incorrect calculations for image depth. *OpenCV* supplies the tools to calculate the distortion parameters as well as the camera matrix [39]. The camera matrix is basically a description for the transformation of points in 3D object space to the 2D image space of the camera. For the used Cameras, we can consider a central projection of points in space onto our image plane. The center of the camera is considered the center of an euclidean coordinate system and the projection plane z is located at distance f equal to the focal length of the camera.

read
and re-
format

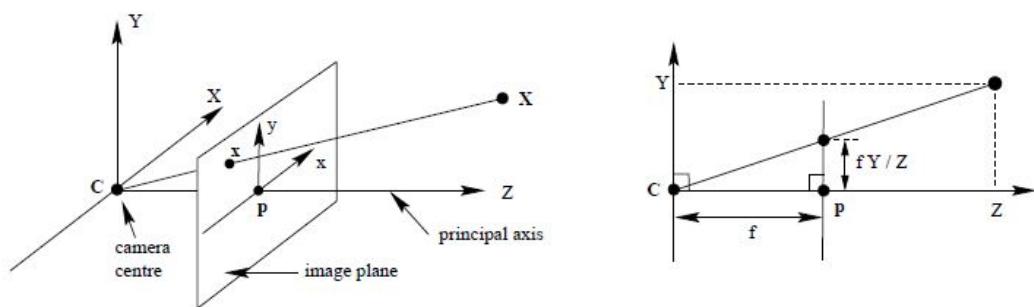


Figure 6.4: Image point projection for pinhole camera systems[18]

As shown in Figure 6.4, a point from the object space can be projected onto the image plane by drawing a ray from the object space point to the camera center. The intersection point of the ray with the image plane defines the projection point. With this knowledge the projection can be described as :

$$(X, Y, Z)^T \mapsto (fX/Z, fY/Z)^T \quad (6.1)$$

For the calculation of the image distortion , OpenCV uses a chessboard image for calibration. The printed image is held in front of the camera at a constant distance and rotated and positioned differently. After every position/rotation change an image of the pattern should be taken. These images are then used to find the corners at the intersection of the black and white squares. With the previous knowledge of the square sizes, the projection errors in radial and tangential directions can be calculated:

[elaborate](#)

[add
Source
code](#)

$$x_{tanDistorted} = x(1 + k_1r^2 + k_2r^4 + k_3 + r^6)$$

$$y_{tanDistorted} = y(1 + k_1r^2 + k_2r^4 + k_3 + r^6)$$

(6.2)

$$x_{radDistorted} = x + [2p_1xy + p_2(r^2 + 2x^2)]$$

$$y_{radDistorted} = x + [p_1(r^2 + 2y^2) + 2p_2xy]$$

As can bee seen from the above equations, the parameters that need to be calculated are k_1, k_2, p_1, p_2, k_3 .

[continue](#)

The python program prototype that is used for tracking the specified color markers is comprised of the following components:

- Image acquisition from camera
- Image conversion to HSV colorspace
- Mask construction and filtering
- Contour finding an position calculation
- Sending of positional data via network as UPD package

6.3.1 Image acquisition from camera

As the computation times for image filtering and mask generation may vary, it makes sense to separate the image acquisition from the computation part. Therefore the loading of the image data frame from the camera is outsourced into its own thread. The synchronicity of the two frame reading threads is assumed for the first implementation and has to be tested on a finished setup.

The tread takes the data of the current image frame and hands it over to the main thread, where the image processing is handled. For the first frames, the full image data is needed

6 System conception

as the position of the markers is not yet known. Performance-wise, this introduces a lot of overhead, since each mask generation step for the specific colors has to go through the complete image data, meaning every pixel has to be read out at least 5 times to get the tracking for all different markers.

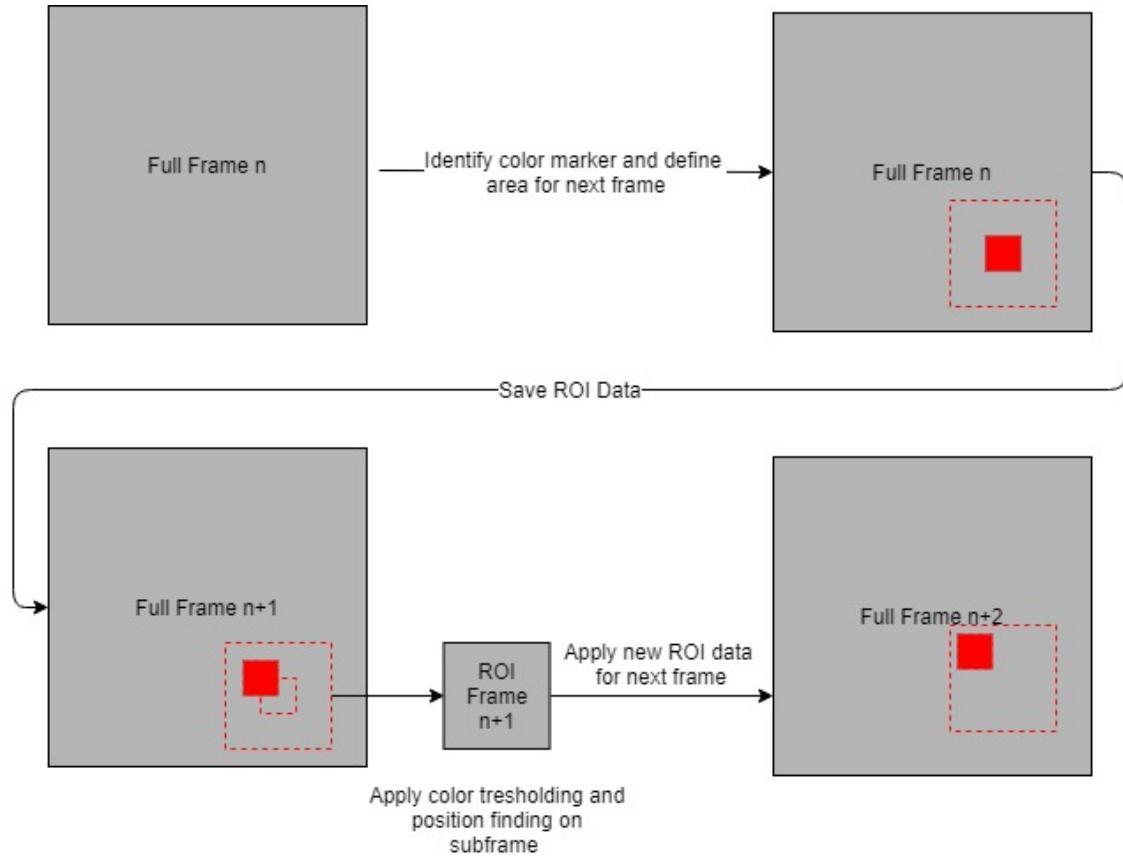


Figure 6.5: ROI calculation for consecutive frames

After a few frames, we get hold of the marker position and can define a reduced region of interest (ROI) on the image (Figure 6.5). The new ROI results from the positional data of the previous frame and has to be dimensioned correctly as to incorporate the possibility of large position differences between consecutive frames. Furthermore the offset position, preferably the top left corner of the region, has to be kept track of. By taking a sub-region of the frame as input for further processing, the resulting coordinates relate to the top left corner of this sub-region as origin. Consequently only the first set of coordinates has the correct origin as reference point for the next frame. Consecutive frames need this information to retrieve the correct sub-region from the full frame by adding the offset to the calculated positional data. First frame ROI calculation is based on a minimum axis-aligned bounding box(AABB) fitted onto the marker. The offset value for the ROI is added to the bounding box values resulting in a rectangle data-set containing the position of the offset top-left corner **tl** and the offset **width** and **height**

of the ROI rectangle:

$$\begin{aligned}\vec{tl}_0 &= (AABB_{x_0} - \text{offset}_x, AABB_{y_0} - \text{offset}_y) \\ \text{width} &= AABB_{w_0} + (2 \cdot \text{offset}_x) \\ \text{height} &= AABB_{h_0} + (2 \cdot \text{offset}_y)\end{aligned}\tag{6.3}$$

These values are taken to extract the sub-region of the consecutive at \vec{tl} with size of width and height. The \vec{tl} value is also saved as offset value OffPos . The position calculation for the proximate frame will utilize the offset value as follows:

$$\begin{aligned}\vec{tl}_{1x} &= x_1 + \text{OffPos}_x - \text{offset}_x \\ \vec{tl}_{1y} &= y_1 + \text{OffPos}_y - \text{offset}_y \\ \text{width} &= AABB_{w_1} + (2 \cdot \text{offset}_x) \\ \text{height} &= AABB_{h_1} + (2 \cdot \text{offset}_y)\end{aligned}\tag{6.4}$$

This should result in the correct positional data for all frames following the initial frame. Finally, the resulting rectangle has to be fitted onto the size boundaries of the full image frame. When applying a fixed offset value, the resulting region that will be extracted from the following frame might reach outside the size boundaries of the full frame. This can happen when the markers are positioned near the edges of the frame. Trying to read out pixels outside the available size of the image will without doubt result into an error by reason of unavailable data at these points. Therefore the combination of position width, height and offset has to be clamped to fit into the available image data:

$$\begin{aligned}\text{width} &= (\max(0, \min(\text{width}, \text{image}_\text{width}))) \\ \text{height} &= (\max(0, \min(\text{height}, \text{image}_\text{height})))\end{aligned}\tag{6.5}$$

Another option which could be evaluated instead of clamping the area would be shifting the offset point to a position where it does not violate the image boundaries and thereby preserving the size of the ROI. Should no marker be found in the defined ROI or the certainty of the result is not high enough, the frame has to be dropped and the next frame should utilize the whole image data.

6.3.2 Image conversion to HSV colorspace

The image data that is supplied by the camera comes in an RGB data format, which we could already use for the further calculation. It does though make more sense to convert the input data into the HSV colorspace. Since we are not using high precision cameras, it is necessary to define a range of color values around the desired color which we want to track. The HSV colorspace is displayed as a cone, in comparison to the RGB colorspace, which is displayed as a cube. The color values for the HSV space are all located on a circle spanning from 0 to 360 degrees. The Hue value (H) corresponds to the angle on the circle, where 0° corresponds to a reddish color, 120° lies in the area of green and 240°

and above correspond to blue. The saturation value (s) corresponds to the amount of white the color contains where 0 is pure white and 100 corresponds to the fully saturated color. For optimal results, only highly saturated colors should be used to ensure correct color detection. The last component is the value component (V) which describes the intensity of the color. Alike the value settings for the saturation, value ranges of at least 50 should be used for tracking precision. For tracking the five fingers of the hand we need five distinguishable colors. Here the primary colors red, green and blue will be the choice for the first three colors. The other two selected colors are orange and yellow. To be able to clearly distinguish these colors in the video frame, a constant and homogeneous lighting is needed as well as a color temperature of the lighting that is in the neutral area to not change the color of the markers.

The color conversion from the input RGB values to the desired HSV colorspace is done as follows:

$$\begin{aligned} hsv_{low} &= (hue_{targetcolor} - sensitivity, 100, 100) \\ hsv_{up} &= (hue_{targetcolor} + sensitivity, 100, 100) \end{aligned} \quad (6.6)$$

6.3.3 Mask construction and filtering

These values are the used as the parameters for a mask generation which generates a binary mask for the current frame where all pixels whose values lie outside of the specified range are set to zero (black) and the remaining are set to 255 (white). For these masks to work properly, any other larger objects that might contain similar colors should be removed from the scene to eliminate wrong tracking data. The used cameras run on an automated white balance and exposure mode. The variation of these values can cause shifts in the appearance of the colored markers and therefore alter the generated masks. To compensate for white balance shifting, a non-white ideally diffuse reflecting background should be used for the tracking space. Also the auto modes should be turned off and appropriate values for white balance and exposure have to be determined at the initialization step of the system. As all digital cameras tend to have signal noise in the sensor data, high frequency noise in the color channels will be present. This noise should be filtered out before any further computation on the image data can be done. The first step in this progress is to use a Gaussian filter to blur the mask.

orig image and mask images for display

The Gaussian filter acts as a low pass filter for the image. The application of the filter cuts out high frequency noise generated by the sensor electronic and other factors. The downside of applying a Blur onto the generated image is a loss of detail. As the markers will not be utilizing complex geometric forms or fine patterns, the loss of details is acceptable.

After this step an erosion and a dilation is applied to the image to further eliminate unwanted noise. A combination of these two morphological operation is used to further improve the data of the thresholded binary image[8, chapter 3.11-12]. Erosion and

dilation operation on thresholded images perform the tasks of removing light spots on dark backgrounds as well as dark spots on light backgrounds. These spots are usually the result of the remaining image noise after the blurring or incorrect pixel values from the camera which cause the pixel to be incorrectly thresholded. To apply the specific operators to the image a mask (often also called kernel) needs to be defined. These masks normally consist of an $n \times n$ shaped matrix where the values in the matrix rows and columns are either one or zero for binary images. An odd number of rows and columns is usually used to be able to define a center pixel. The central entry of the matrix corresponds to the current pixel of the image for which the morphological operation is to be applied. As the data of the pixel should be preserved, the value is set to one.

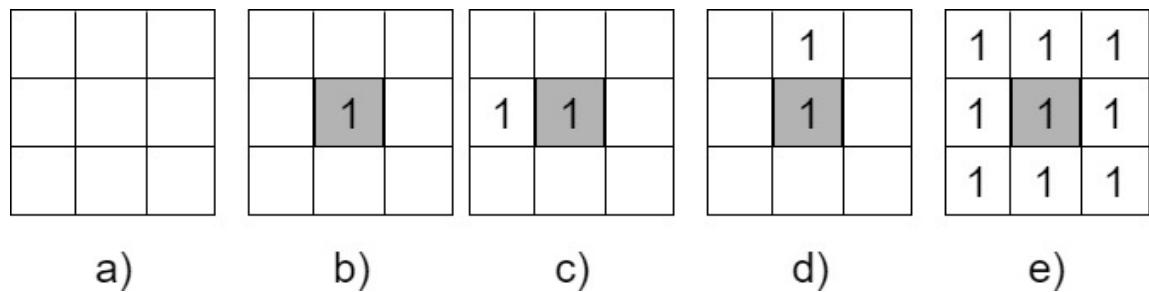


Figure 6.6: Example Mask of 3x3 size :a) empty mask b) Mask that keeps pixel value c) combination of b) with a left shift d) combination of b) with an upshift e) isotropic mask

The processing of these matrices is not very complex. The central pixel value and the values of the neighboring pixels are taken and summed up. The ones in the matrix outside of the center value define if the pixel is taken in for the operation. A threshold value is then defined. The result of the sum operation is compared to the threshold value. If above the threshold, the central pixel value is altered, depending on the image operation. If below, the current value is kept.

When using for example the isotropic matrix displayed in Figure 6.6 e) and setting a threshold of 8 on a dilation operation, the result would be the filling of a white pixel when surrounded by black pixels. The same operation on an erosion would lead to an removal of a black pixel surrounded by white pixels.

The combination of the morphological operator erode and dilate leads to two methods called *morphological opening* and *morphological closing*. The idea behind these two is to use the two basic operations in a certain order to enhance the binary image quality [8,

input: threshold $threshVal$ value for the calculation, image segment
 $imgSegment$ of size $n \times n$, mask $morphMask$ of size $n \times n$, operation
 value (0 or 1) $opVal$, value of central pixel $centVal$

Result: New value for pixel

```

for  $i \leftarrow 0$  to  $maskHeight$  do
  for  $j \leftarrow 0$  to  $maskWidth$  do
    if  $morphMask[i, j] = 1$  then
      |  $result \leftarrow result + morphMask[i, j]$ 
    end
  end
end
if  $result \geq threshVal$  then
  |  $centVal = opVal$ 
end

```

Algorithm 1: Pseudocode for the the pixel value calcuation

chapter 3.12]. The formal definition for the operation:

$$\begin{aligned}
 O &= \text{Original image} \\
 \oplus &= \text{Dilation operation} \\
 E &= \text{Erosion mask} \\
 \ominus &= \text{Erosion operation} \\
 D &= \text{Dilation mask} \\
 \circ &= \text{Morphologic opening} \\
 \bullet &= \text{Morphologic closing}
 \end{aligned} \tag{6.7}$$

$$O \bullet D = (O \oplus D) \ominus E$$

$$O \circ D = (O \ominus E) \oplus D$$

The *morphological closing* operation is applied first to eliminate so called "salt noise" (white pixels in black area), narrow cracks, channels and small holes in the original image. The major part of the work for this operation is done on the areas where not tracking marker is found. It removes the remaining noise left over from the blur or signal distortion. The application of the *morphological opening* removes so called "pepper noise", fine hairs and small protrusions. The outcome of the practical application of the opening is the removal of wrong pixels inside the area of the color marker in the image. This helps with the later calculation of the bounding boxes as the algorithm searches for the largest closed area int the image. Generally the closing of an image enlarges it while filling bright defects inside of an black area and the opening makes it smaller while removing bright defects on a dark surface. As the system will not be handling large image data,

the mask sizes can be kept low, utilizing the above described 3×3 mask size. The structure options provided by *OpenCV* contain a cross structure, an elliptical structure and a isotropic version. As it would be useful to get all surrounding image data and the tracking markers will not have complex geometric structures, the isotropic mask structure should be sufficient. The threshold limit should be set to an initially reasonable value with further testing on the running setup.

6.3.4 Contour finding and position calculation

With the cleaned mask we can continue and search for the white areas in the mask which might represent our target. Under the assumption that we have removed all other parasitic objects from the image, the marker should be the largest area of positive pixels in the mask frame. Therefore only the largest area found in the search is taken as the desired tracking marker. For the selected area, a fitting bounding box is calculated and the center of this box is used as the positional parameter of the tracking marker.

6.3.5 object tracking

6.4 Systems communication

The data that is generated on the two Raspberry's has to be communicated to the "parent" unit where further processing is done. The data delay between the different endpoints should be minimal. As the data is streamed in real time, protocols like TCP, which utilize control mechanisms for package order and receiving of data use to much overhead. In the time frame needed for the protocol to request a missing package again, the data will already be obsolete.

Furthermore, the amount of data that is send in each data frame is minimal, as this is only the position value of the specified number of tracking points and some separators for string processing. Since the devices will be connected to a local network via a switch, network workload from other devices and protocols can be assumed as minimal and not impacting transport time.

Regarding all these points a simple UDP protocol setup seems to be the best choice. Package loss that might appear with UDP is not critical as it means only the loss of one frame Data set.

It has to be ensured that the UDP data streams from both devices are recorded separately to ensure that the data can be separated correctly for the stereoscopic calculations. The possibility to send multi- and/or broadcast messages via UDP also open up an easy way of communication between the "parent" and the Pi's for control communication.

add
object
tracking
method
-i vi-
sual or
HTC
vive
tracker

6 System conception

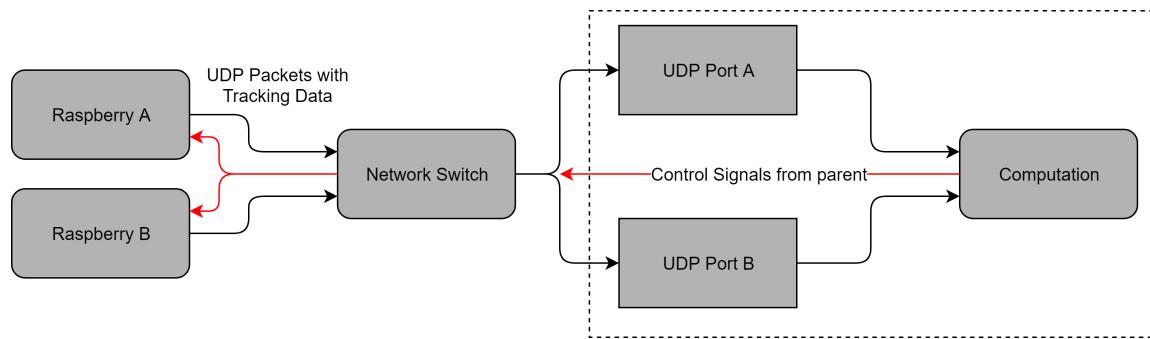


Figure 6.7: Workflow for the UDP Data packages that are send from the Raspberryys to the parent PC

6.5 Data processing on the receiver side

The parent unit receives the two UDP data streams from the Raspberry's. To ensure that the two data streams are separated and therefore distinguishable, each stream receives a dedicated listening port on the client side (Figure 7.7). This separation of data streams also makes it possible to separate the UDP listeners to separate running threads in the background, taking away the task of data processing from the main thread which is already tasked with the rendering of the virtual scene. The positional data will be encoded as comma separated x and y coordinate set. Data separation is done by a simple semicolon as a separator between the coordinate sets. This takes very simple substring operations on the client side to extract the single coordinate sets from the string that is sent with the UDP package. After the string decoding, the extracted values are written into thread-safe variables. This allows the system to read current data without conflicting the writing process of the UDP threads.

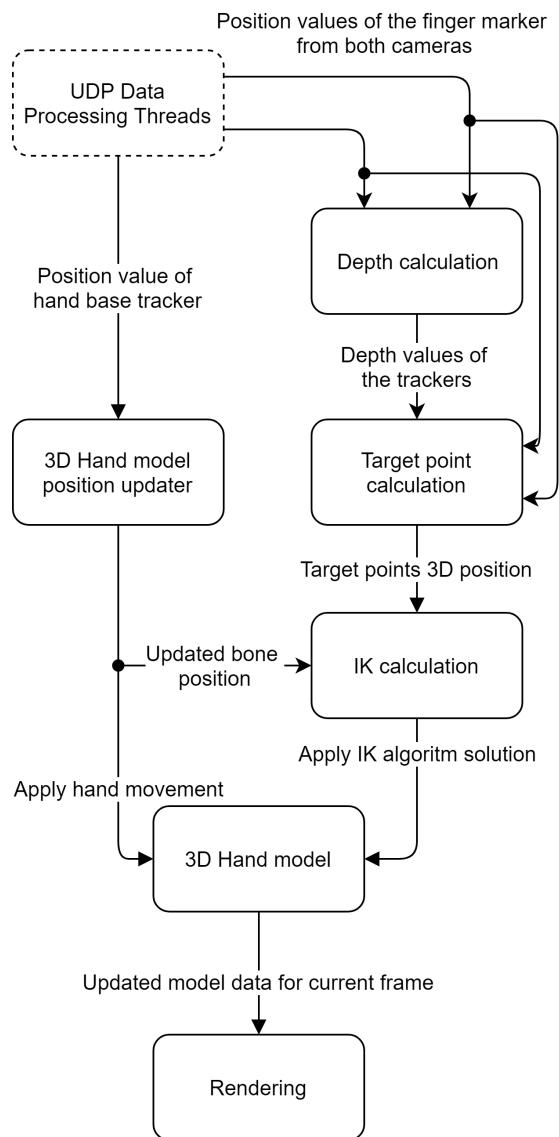


Figure 6.8: Overview diagram of the processing steps for the hand model data on the client side

6.5.1 Stereoscopic calculations

The positional data that is extracted in the UDP data processing only consists of 2D position data of the tracked marker on the image plane. To retrieve depth information for correct spatial placement of the marker the image disparity of the two cameras can be utilized to calculate the object distance from the cameras. [50]

Before diving into the mathematical concepts, some essential terms have to be explained. These are:

- Interaxial distance
- Parallax
- Zero-point plane

Interaxial distance

A stereoscopic imaging system is based off of our own human visual system as it produces stereoscopic images. Therefore such a system always consists of at least two cameras which are mounted in parallel. The distance difference between the two cameras is called the "inter axial distance" of the system. The usual distance for such systems is the mean distance of the human eyes, which corresponds to around 75mm. Shifting the cameras closer together or further apart will change the resulting stereoscopic effect, where shifting the cameras closer together will enlarge objects and shifting the cameras further apart will make objects seem much smaller than they are.

Parallax

Since the cameras are separated by the *interaxial distance*, a disparity in the generated images will occur. This means that the objects in the picture of the right camera will have a different position in terms of perspective than the images that result from the left camera. This resulting disparity is used to generate the 3D effect when displaying the images. The term "*parallax*" is used to describe the comparison of the two images. The two images can have a negative, positive or zero *parallax*.

Zero-point plane

The *zero point plane* is a special theoretical plane in a stereoscopic system. The location of the *zero-point plane* defines the distance from the camera system at which the image points from both cameras match. Objects that lie before the *zero-point plane* will have a negative *parallax* and will appear closer to the viewer. On the opposite side, objects with positive *parallax* will lie behind the zero point plane and therefore will appear further away.

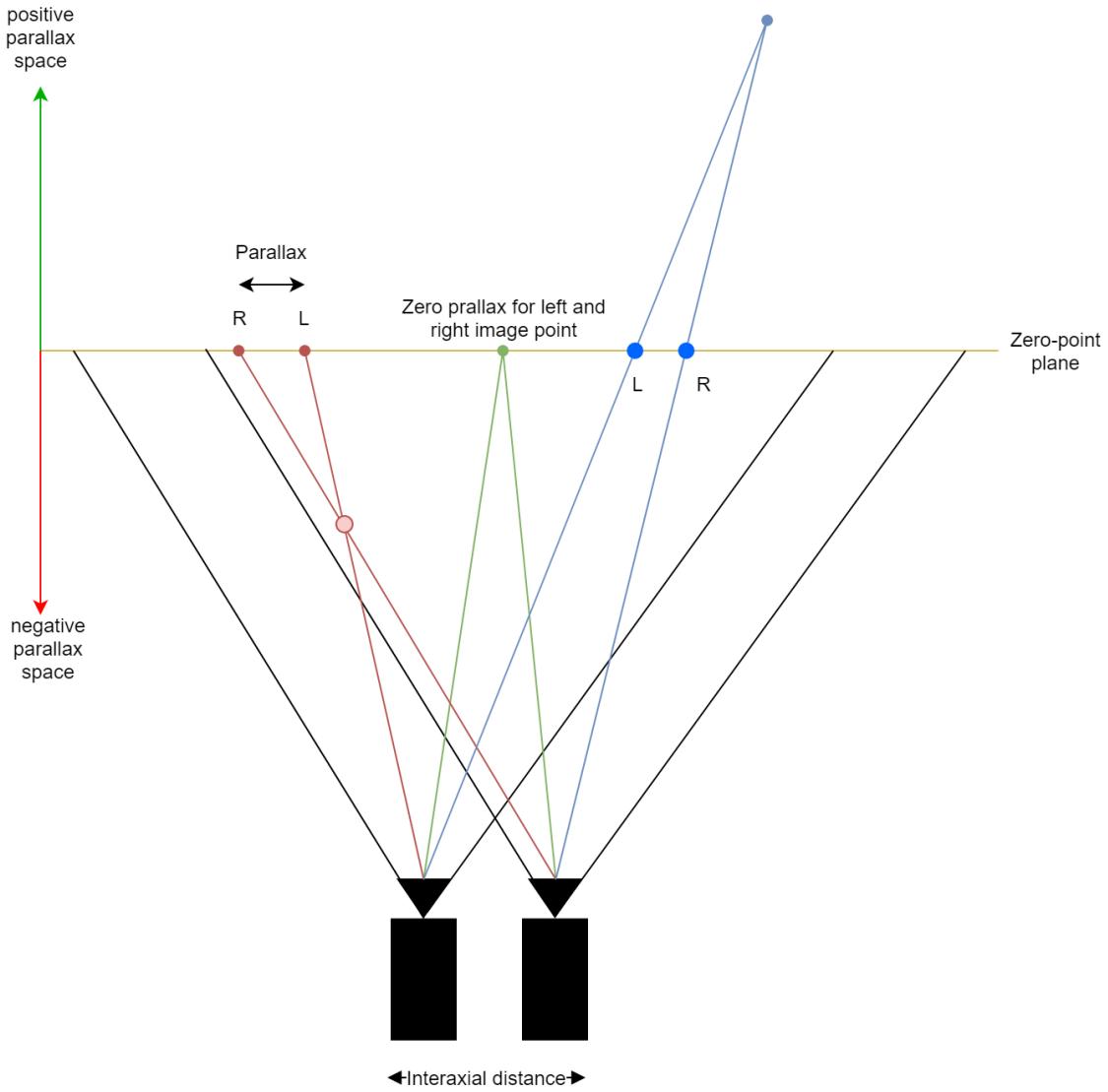


Figure 6.9: Stereoscopic areas

Additional attention needs to be put in the camera setup before calculating depth from disparity. The cameras need to be checked for horizontal alignment to achieve correct values for the calculation.

6.5.2 Depth calculation

With correctly aligned system, one can calculate the distance of an object from the cameras using common trigonometric fundamentals[24, 57].

The position for the right camera will be denoted as S_r and the position of the left camera will be denoted as S_l (see Figure 6.10a). The resulting distance of $S_l - S_r$ is the *interaxial distance* B of the system. The view angle θ of the Camera, which should be the same for both cameras, will respectively be θ_1 and θ_2 . From the camera positioning, the assumption of parallel optical axes is made. The angles of ϕ_l, ϕ_r describe the resulting angle between the optical axis of the camera and the object. By using geometric derivations we can express that:

$$B = B_1 + B_2 = D \tan \varphi_1 + D \tan \varphi_2 \quad (6.8)$$

Rearranging for D gives us:

$$D = \frac{B}{\tan \varphi_1 + \tan \varphi_2} \quad (6.9)$$

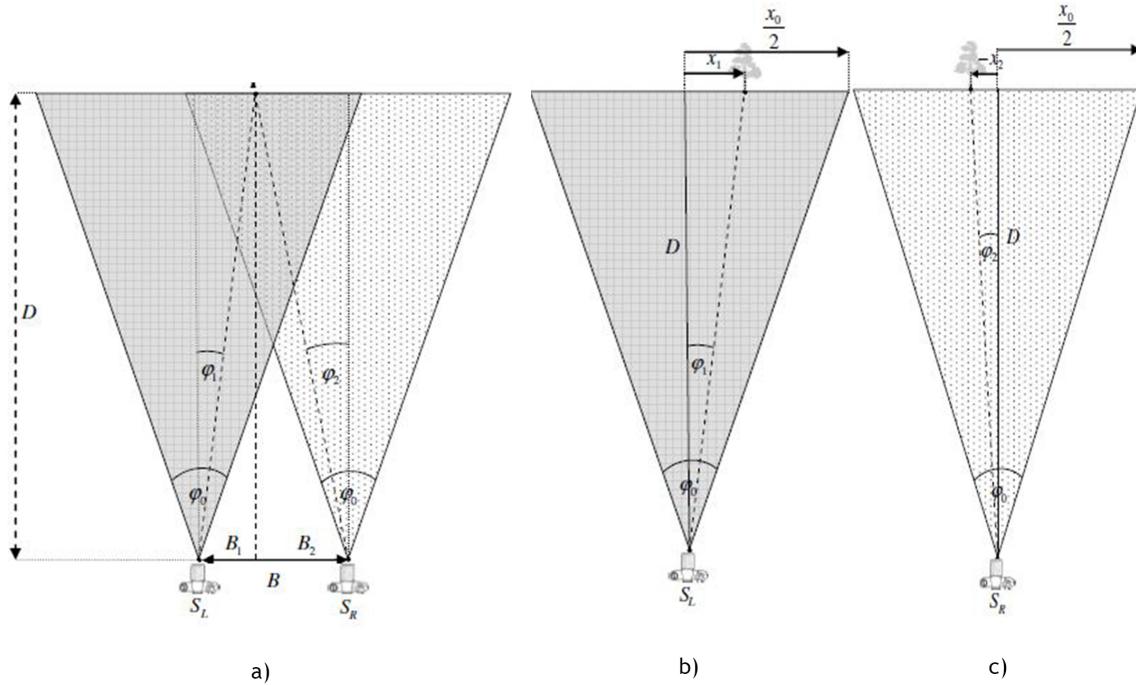


Figure 6.10: Dimensions for calculation. Images from:[24]

With Figure 6.10b-c one can find that:

$$\frac{x_1}{\frac{x_0}{2}} = \frac{\tan \varphi_1}{\tan(\frac{\varphi_0}{2})} \quad (6.10)$$

$$\frac{-x_2}{\frac{x_0}{2}} = \frac{\tan \varphi_2}{\tan(\frac{\varphi_0}{2})} \quad (6.11)$$

Resulting in the calculation of D to be expressible as:

$$D = \frac{Bx_0}{2 \tan(\frac{\varphi_0}{2})(X_L - X_R)} \quad (6.12)$$

Where B is the distance between the cameras, x_0 being the horizontal width of the picture format in pixels, φ_0 representing the viewing angle of the camera and $(X_L - X_R)$ representing the disparity between the two images in Pixels.

6.5.3 Model position update

Before a solution with the inverse kinematics algorithm can be calculated, the hand movement in real space has to be matched with that on the digital side. To achieve this, a further marker besides the finger markings is required. A sixth color marker has to be positioned at the connection point between hand and forearm to serve as a global position reference. The position and rotation difference between the last and current frame can then be calculated from the tracking values and then be applied to the whole model.

6.5.4 Inverse Kinematics

For the inverse kinematics calculation, the *Fabrik* algorithm that is described in Section 2.2.7 was selected. It does not require complex matrix calculations and has a fast convergence rate. For the implementation of the algorithm on the client side, the *Ca-liko* Java framework[29] is used. There are also frameworks for other languages, but the java framework is platform independent. The library supplies all the needed classes and functions to create kinematic chains and structures as well as the implementation of the solving algorithm. It furthermore ships a visualization package which utilizes the Lightweight Java Game Library 3 (LWJG3). These classes can be used optionally for debug visualization or even as a base for a more sophisticated visualization as it is possible to load further components for i.e. VR into LWJGL.

For the inverse kinematics calculation to work properly the updated bone positions and the calculated target points from the finger tracking are needed. For each finger a kinematic chain is set up with the correct bone length. The length of each bone has to be measured manually for the first prototype. Further iterations should provide a more elaborated and maybe even automated calibration procedure. The movement of each bone in the kinematic chains has to be restricted by constraints to restrict the algorithms solutions to physiological possible hand poses. The values for finger constraints that are described in Section 2.1 should provide sufficient values. The resulting kinematic chains all have their own target point for which the algorithm can solve for. All chains will be grouped into a kinematic model for easier data handling. The kinematic model has a base and a connection structure through which the fingers are connected to it. With this data, the algorithm is capable of calculating a solution for each kinematic finger chain and applying these values to the position. If the target is reachable, the digital finger position should match the pose of its real world counterpart.

7 System Evaluation

7.1 Camera hardware and control

The image processing of the camera frames on the raspberry can be a significant bottleneck for the system. Real time evaluation demands the processing time of a frame to be in the range of the time of the camera FPS. With the current system setup running at 30fps, this means that the image processing time, containing frame readout and downstream processing of the images, has to be done in about $1/30\text{s}$ or 33 ms to achieve "real time" processing. Professional stereoscopic setups use cameras with global shutters and a synchronization signal. This enables a sub frame synchronization of the two camera systems. The cameras used in the prototype setup use a rolling shutter instead. Since these cameras are designed to be of easy use for the normal consumer, they lack means of external synchronization. Furthermore, electrical drawings as well as controlling software are closed source. To achieve a form of hardware synchronization, a large amount of reverse engineering would be needed. The multi layered circuit board of the camera would also complicate the reverse engineering step. As the driver software is closed source, means of getting system information from and to the camera hardware for software synchronization is severely limited. For these reasons, attempting the synchronization of the data further down the processing pipeline is more reasonable.

To synchronize the reading start point of both cameras the system boots up to the point where all needed components are initialized. At this point the program waits for a start message from the parent system via UDP. The parent system send the start message at the end of its initialization phase via a UDP broadcast, ensuing that both Raspberry's get the message at the same time.

In the normal usage mode, the camera uses mostly automatic modes for parameter setup. This continuous measuring mode showed fluctuation in image brightness and color saturation. To eliminate this problem, all automatic modes are turned off in the camera initialization phase and fixed values are loaded from a JSON file. To get the values for the JSON file a calibration step was implemented into the command line tool. This calibration step enables the user to preview and manipulate the following values for the camera:

- Saturation
- Brightness
- Gain
- Exposure time
- Contrast
- White balance values for blue and red channel

The user set calibration values can be written into a new JSON file and are directly loaded into the program after calibration is finished. Control over these parameters is crucial for achieving a constant color separation.

The threshold values values for the color tracking also need to be initially calibrated. These values are dependent on the lighting situation and any change results in the need for re-calibration. To simplify this calibration a calibration function in the same fashion as the camera calibration was implemented. It prompts the user a preview of the resulting thresholded black and white image with applied camera settings. Range sliders for lower and upper boundaries for the three RGB components make it fairly easy to adjust and view thresholding results in real time. The calibration values can also be stored into a JSON file. The calibration values will be fed into the system after finishing and can be saved to a JSON file. This File can be read out on following program startups to eliminate the need of calibration at each startup.

image
of cam-
era and
color
calibra-
tion

7.2 Image processing performance

The first prototype implementation with *OpenCV* in *Python* showed, that a speed of 33 ms was reachable when only tracking one color. The implementation reached around 30-40 ms processing time when scanning whole images every time. An implementation of an "*Region of interest*" feature broke down the processing time to around 30 ms for a single color. The downside of the implementation was revealed when implementing the other 4 colors needed to do a full five finger tracking. This approach ramped-up processing time to around 100 ms per frame, making this solution run at 10 FPS.

The sequential algorithm also showed another design flaw. When not utilizing the ROI approach, the algorithm would scan the whole grabbed camera frame for each color to create the corresponding threshold map. The *OpenCV* color threshold methods are designed to only search for one color threshold at the time, therefore needing this procedure. One solution option is the already mentioned ROI usage. This solution would furthermore benefit from a multi threaded implementation, as we are not manipulating the original image. Parallel memory readout is not a problem and the generated mask can be saved separately.

Another approach which could help speed up the process is implementing an own method for thresholding the grabbed frame with the possibility to do all the needed color thresholds in one image loop. With this solution, the overhead would be reduced by 4 image iterations, thresholding operations would stay the same as these still have to be run on each pixel. The output of this method would then return the 5 needed threshold mask for further processing.

A second performance issue that arose from the first prototype was that the input camera frames came RGB coded. For better options of color separations, the image was initially converted into HSV colorspace. This operation turned out to be second longest operation after frame grabbing.

Testing of the cameras showed that the color segmentation in the original RGB images of the cameras are good enough for the used test colors, causing this step to be omitted and shaving off several milliseconds of processing time. The rest of the processing steps were lying in the sub millisecond range and are therefore not as valuable for performance optimization. The results from this prototype showed, that the *OpenCV-Python* combination is not suitable for reaching the 30 ms target time. *Python* is not a very performant language in this area. The OpenCV library used by python is actually just the ported version of the C code library with bindings for python. As C and C++ are more hardware near and therefore more performant, the second prototype implementation was done in C++.

The ported version of the Python code to C++ initially showed similar performance measurements when using the sequential approach. This was to be expected, as it is the same code the python bindings are using. Further investigation into code timings showed, that another bottleneck was the image optimization feature which applied an erosion and dilation to remove high frequency noise in the image. This operation brought the time up to 100 ms processing time when processing the whole frame area, making the algorithm not usable for "real time" application.

Removing this feature brought a major speed up in processing time. Furthermore, the implementation of a thread pool, which handles all the color detection jobs, brought down the calculation time for 5 colors to around 120-140 ms. Although these times are still not usable for real time, it brought a significant performance boost.

A test with lower resolution than 640x480 showed that the calculation time can be reduced

Add
mea-
sure-
ment
values

furthermore. This indicated that the ROI idea would be feasible for further performance optimization. A test at a frame resolution of 320x240 pixels reduced the processing time for all five colors to below 10 ms.

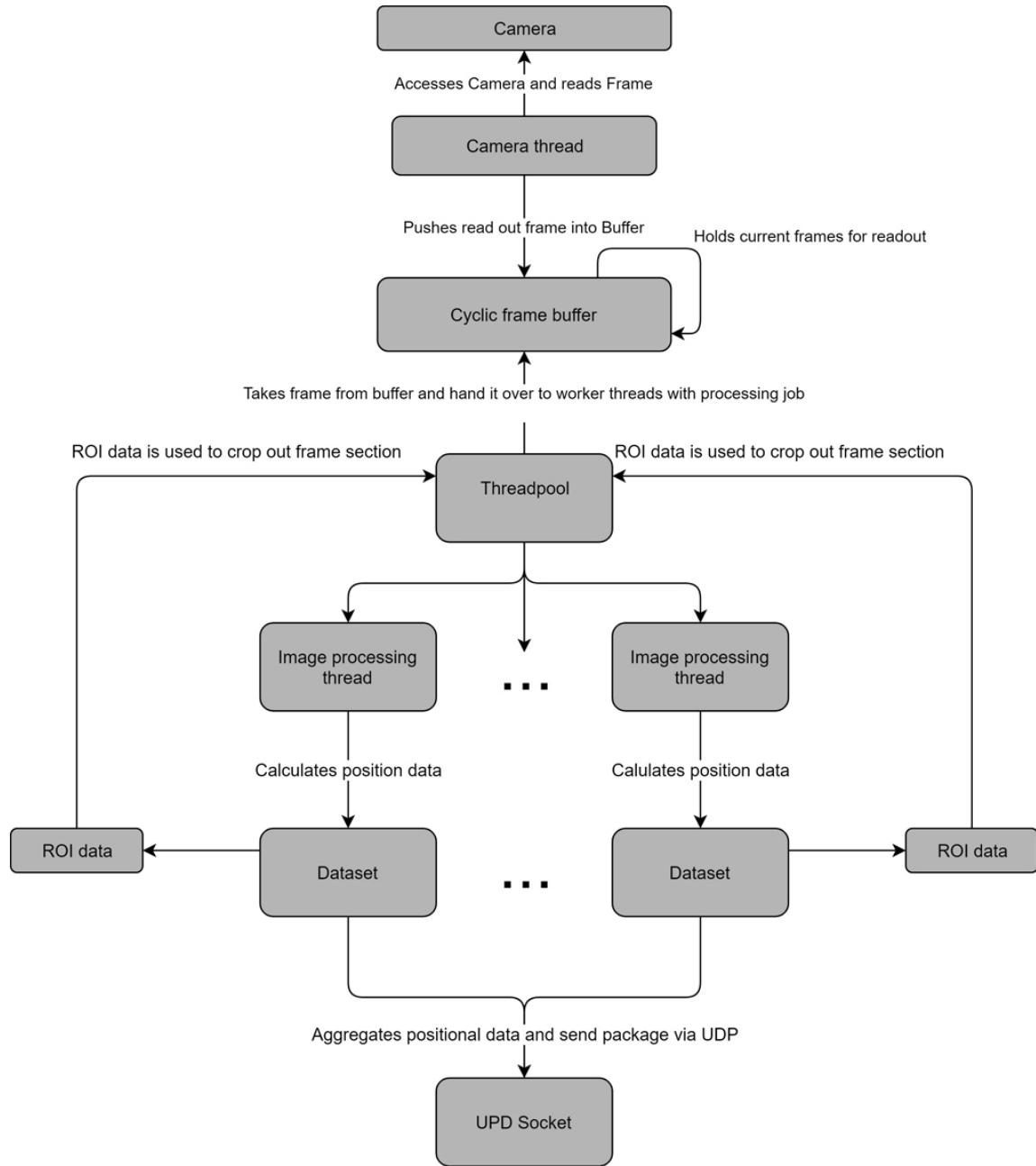


Figure 7.1: Work flow of the C++ implementation with thread pool solution and ROI

Utilization of the knowledge from these tests led to the implementation of the aforementioned "region of interest" feature (see Figure 7.1). The first frame is analyzed as a whole frame, optimally resulting in the detection of a marker. As the marker detection creates a rectangle around the tracked area, we already have the coordinates for the ROI and only have to add an offset to this area.

The value of the offset has to be set high enough to ensure that in the following frame,

the tracking marker is still contained in the ROI. It also has to be ensured, that the generated ROI is clamped to the frame dimensions. If not doing so, parts of the readout are located outside of the image plane, causing readout errors when used for the subsequent frame. The following frames will then be processed with the input of the ROI, selecting only the defined section of the image and updating the ROI with the new data for that frame. Performance measurement for tracking consistency with stationary color trackers was performed for a time period of 10.000 frames to get qualitative results. The sequential implementation showed an average processing time of 46 ms with a standard deviation of 11 ms.

Further analysis showed that the high standard deviation came from peaks in performance times that took up to 400 ms when loosing tracking. Filtering all values over 50 ms out from the data set showed that only 1% (~ 1000 out of 10.000 frames) of the recorded processing times were lying outside of this time frame. Filtering out these values from the data set and recalculation dropped the average processing time down 3 ms to around 43 ms. The standard deviation was dropped drastically from its previous 11 ms to around 1.6 ms. This shows that a filtering of processing times on the Raspberry would lead to a more consistent frame processing time. To do so, after every processing cycle the time needed is checked and if it is over the defined time limit, no data is send.

Since this performance measurement does not show which part of the process consumes large amounts of processing time, the same measurement was done with a finer time tracking on the single parts of the processing pipeline.

Table 7.1: measurement results for sequential processing timing measurement

Time in ms	Image retrieval	Stereo rectification	Color detection (6 colors)	UDP sending	Summed time
Average	0.91685	31.39436	14.55122	0.12053	46.96780
Std.Deviation	0.17105	7.19069	5.06205	0.06031	10.77171

Table 7.1 highlights the performance bottlenecks of the system. Image retrieval from the camera as well as the time to process and send the data results via UDP are in the sub millisecond range, therefore they impose no significant delay. Without these two steps, the remaining 96% of the processing time is taken up by the image stereo rectification and the color tracking. The processing time of the color tracking could be brought down further by the use of multiple threads for processing, but as the comparison values show. The far larger time consuming part is taken up by the stereo rectification part with nearly two thirds of the processing time. The stereo rectification is needed for correct disparity measurements.

measure
multi
thread-
ing

rectify
maps
con-
version
and

7.2.1 Color accuracy and ROI size

As already mentioned in the conceptional phase, the lighting of the tracking area is a major factor. Variations of lighting intensity and color temperature cause the color of the trackers to shift their color. This can cause a fluctuation in the calculated tracker positions, as the defined color threshold ranges are kept as small as possible to achieve a clean separation of the tracker colors and reduce unwanted noise.

System testing with the 5 different colored shrink tubing markers showed, that the color tracking for colors outside of the color range of the human skin tones (green, blue) is rather unproblematic. The color markers falling into the colorspace of possible light reflection from skin may cause problems, depending on the lighting situation of the setup.

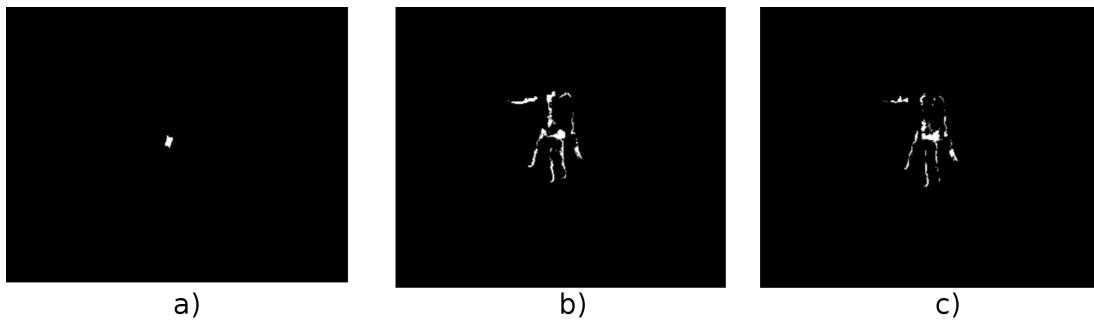


Figure 7.2: Images showing reflection problems with orange color thresholding. a) Correct threshold with only the marker being thresholded b-c) Depending on the rotation of the hand, skin reflection color falls into tracking threshold.

Figure 7.2 displays such a case where depending on the orientation of the hand, reflection from the scenery lighting can fall into the color tone space of the tracked marker, causing large parts of the hand to light up in the image. Since the calculation of the tracking marker position relies on finding the largest closed "white" area in the thresholded image, this can cause a "jumping" of the calculated position.

The first solution for this problem is to narrow down the color threshold spans. By doing so, one can ensure that only the wanted colors are tracked. This method does have a downside effect. When the markers are moved in the tracking space, light variations and shadowing will cause the colors to shift their appearance slightly. With broader threshold ranges, this does not seem to be a large problem as these changes still lie inside the ranges. Narrowed down color ranges might not include these color variations, creating cavities in the thresholded image.

The applied morphological operations on the image can fix these problems up to a certain point, depending on filter kernel size, but for larger areas this could vary position tracking results.

The implemented ROI feature speed up the whole system calculations once the colors are found. Before this point, the system scans whole frames to find colors, which takes more time than the much smaller ROI regions. Empirical evaluations showed, that

for the used tracking space, a ROI region size offset of 40 px in x and y directions produces the best results in terms of consistency. Smaller region offsets caused the system to use tracking for faster hand motion, which causes system slow down until the tracking has recovered. Generally, larger ROI's produce more constant tracking results at the cost of higher calculation time.

7.2.2 Finger markers

The initial color marker size was selected to be relatively large, spanning most of the last segment size of each finger. The used material of the colored shrink tubing was selected because it seemed to be relatively diffuse and therefore unprone to errors resulting from the reflection of the scene lighting. The material turned out to be mostly diffuse reflecting, but depending on color and angle of light incidence, reflections in the color of the scene light did occur. A test set of smaller marker was also evaluated, showing similar tracking results in comparison to the larger markers and also the same color reflection problems.

The benefit of a smaller marker size would be much lower restriction of the finger movements and also unveiling the fingertips of the fingers for better haptic interaction. As both markers showed the above described problem of mixing marker area with possible skin reflections, and improvement to the markers was applied. At both ends of the larger markers, a piece of gray colored tape was applied to create an artificial border for the tracking algorithm.



Figure 7.3: First set of shrink tube color markers

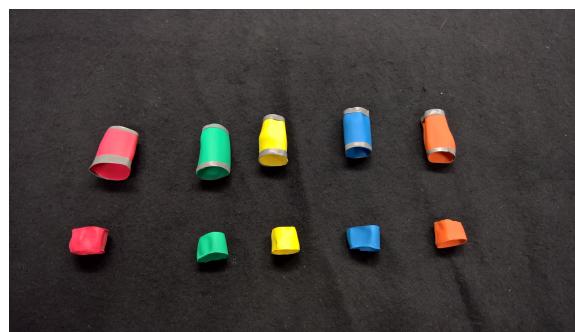


Figure 7.4: Improved larger and smaller set of shrink tube color markers

The idea behind this solution was to create a fixed border between the color area of the tracker and the color area of the skin tones. The grey color of the border ensures that this area will not be tracked. The usage of black instead of gray would even furthermore ensure this.

Since the threshold algorithm searches for the largest closed color area in the binary

threshold image, this border should cause a separation of unwanted skin reflection area and actual tracker color area. The problem that the resulting color marker area might not be the largest area still persists after this improvement, but this case has to be handled separately.

Optical results showed that for the colors lying outside of the red to orange color spectrum that skin reflections produce, the artificial border stabilized the color detection.

The idea of using shrink tubing as marker material showed to be suitable except in the cases described above. One downside of the material is that the displayed colors are already the highest variety of available colors on the market. Other colors from the green and blue color areas like purple are simply not available or have to be ordered as a special request in larger amounts.

Also the original diameter is much larger than a human finger. The used shrink tube is able to be shrunk to one third of its size. The reason for this was to be able to adapt to the varying diameters of the human hand. The shrinking procedure requires a constant amount of heat to be applied to the material to activate the shrinking. The amount of heat that is needed surpasses the heat production of a standard hair dryer since the original application of shrink tubing is on heat resistant cables. The heat that needs to be applied can easily be more than 100 degree Celsius, which makes a direct application on the users hand highly insecure.

For the prototype, the shrink tube parts were shrunk with a standard cigarette lighter and continuously fitted onto the user finger until the desired form was reached. This method showed to be sub optimal, since the flame of the lighter produces only a punctual heat source. This caused uneven shrinkage of the heated parts. This can lead to cavities or bumps in the material and change the reflection properties at these points. A regulated heat gun would probably generate better results.

To get smoother shrinking results, parts of PVC tubing with the correct diameter could be used as dummy parts for the main part of the shrinking.

Another approach that was tested was the usage of matte acrylic paint to cover parts or even the whole finger. The matte acrylic paint comes with the benefits of being easily to apply to the finger. Acrylic paint is available in much more color variations than the shrink tube. The finger coating is dry in under a minute after application. Adaption for finger size is automatically included in the application process. If the tracking system does not need to evaluate for hand rotation, i.e. the working principle is a top on view, a nail polish style application of the color on the fingernail can be sufficient for the tracking system. The thumb needs to be treated separately, as it has more rotation capabilities than the other fingers. The thumb should be marked completely on all sides to achieve a constant tracking. If hand orientation is tracked, the other fingers should be completely marked as well.

replace
with
image
from
final

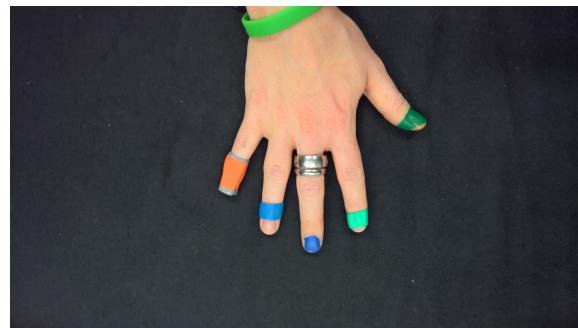


Figure 7.5: Different types of tested markers. Thumb with complete acrylic coloring, index and ring finger with smaller shrink tube markers, middle finger with nail polish style acrylic color and little finger with large improved marker

The best tracking results in terms of stability and color separation were achieved with the acrylic paint as marker material. For the colors a mixture of blue and green color tones together with red tones in the purple and magenta section were chosen as the final colors. The usage of the acrylic color also provides the least degradation in terms of gripping haptics on the fingers.

Measurements

7.2.3 Depth measurement accuracy

To determine the accuracy of the system for its depth measurement values a simple test bench setup was used. The camera rig was aligned horizontally and fixed to the test bench. A large sized colored marker (75 mm x 50 mm) was used as target for detection. The marker was positioned at altering distances from the camera rig along the central axis of the camera rig. The height at which the camera rig is positioned in the experiment setup will be around 100 cm, so the measured distances started at 100 cm from the camera and were decremented in steps of 5 cm until 20 cm in front of the Camera Rig(see Appendix Table 10.1 and Figure 7.6).

7 System Evaluation

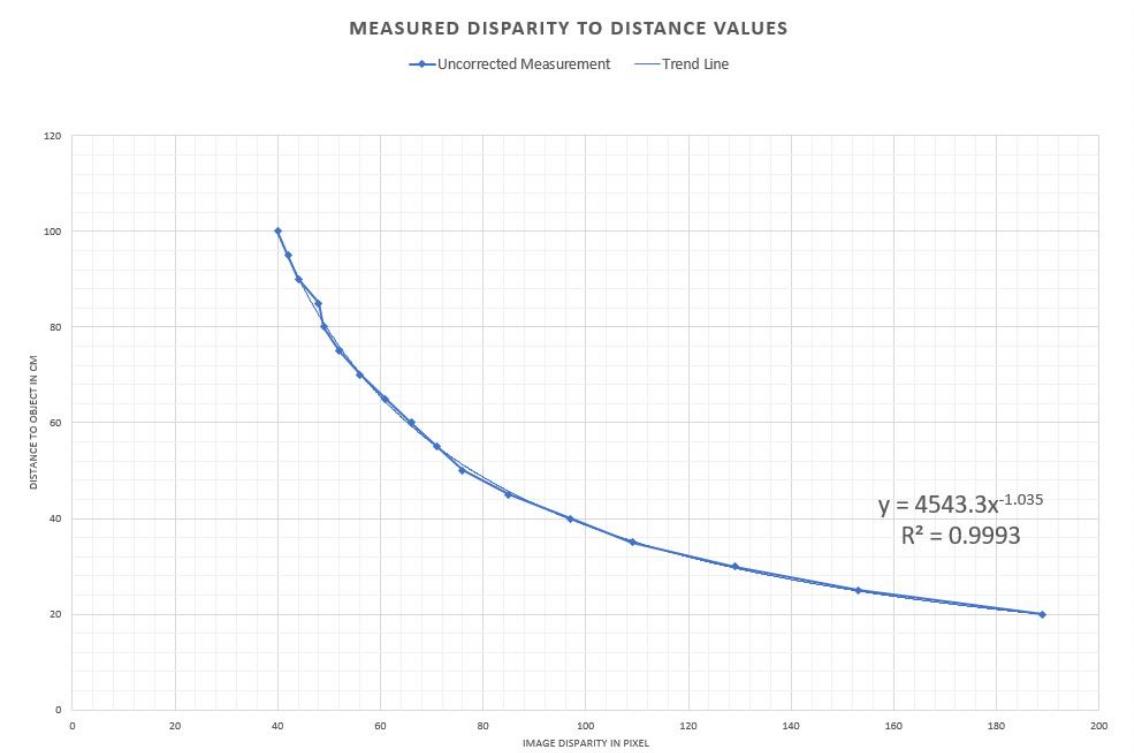


Figure 7.6: Chart showing the results of the disparity measurement with a plotted trend line

The accuracy of the camera system is limited by the number of pixels in relation to the camera view angle:

$$\Delta\varphi = \frac{\varphi_0}{x_0} \quad (7.1)$$

With the system set up at 640 pixels of image width and a horizontal viewing angle of 62.5° , $\Delta\varphi$ is equal to $0.0977 \frac{1^\circ}{pixel}$. The resulting error would be:

$$\frac{\tan \varphi}{\tan(\varphi - \Delta\varphi)} = \frac{\Delta D + D}{D} \quad (7.2)$$

$$\Delta D = \frac{D^2}{B} \tan(\Delta\varphi) \quad (7.3)$$

rework
text
formulation

The system error for a $D = 100\text{cm}$ would therefore result in about 2cm of possible error. As the measured values show deviations of up to 5% from the correct value, the function from the trend line displayed in Figure 7.6 was taken as a correcting factor for the depth calculation.

Equation 7.3 uses the uncorrected values for the calculation. Values lying on the trend line of Figure 7.6 should correspond more exactly to the correct distance values [33]. The equation can be rewritten to:

$$D = k * x^z \quad (7.4)$$

with K being:

$$k = \frac{Bx_0}{2 \tan(\frac{\varphi_0}{2} + \phi)} \quad (7.5)$$

and x the disparity in pixels. The ϕ term in the equation above is used as a compensation for possible alignment errors. The trend line, which is fit to the measurements in Figure 7.6 represents the function needed to fulfill Equation 7.4. The calculated values are $k = 4543.3$ and $z = -1.035$. With the utilization of the corrected value formula, the accuracy of the depth measurement results is acceptable for the prototype application. For future work, the distance measurement procedure could be applied with more measurement points to refine the resulting function for more precision. The measurement values that were retrieved can be found in the Appendix Table 10.1.

7.3 Inverse kinematics algorithm

- missing feature for parabolic constraint curves
- measure calculation times
- setup difficulty

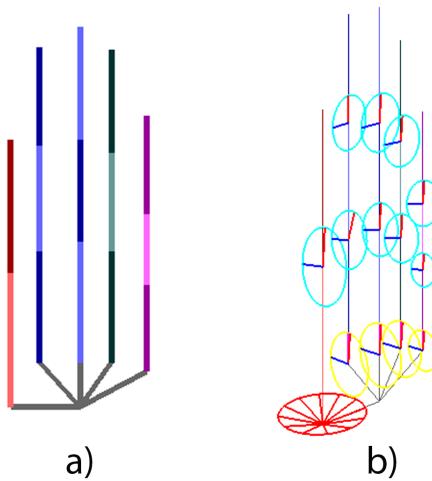


Figure 7.7: Hand model representatio of the used IK Framework: a) Kinematic chains and base structure, b) hand model with visualized movement constraints

8 Future Work

The system Evaluation showed that the principal functionality of the tracking system can be established with the selected component. One major point of improvement that could be applied to the system would be improvements to the used camera control framework. At the current state, a max frame rate of only 30 FPS readout is reachable while the hardware is theoretically able to supply up to 90 FPS readout. Large parts of the processing time on the raspberry side are accounted for by costly image optimization like blurring or stereoscopic transformations. These operations on whole images are classical candidates to be outsourced onto the GPU of the Raspberry for faster computation. A conversion of the existing OpenCV code into the Assembler like code for the raspberry GPU should be a future goal. -improve depth tracking value precision through a more elaborate calibration session where the depth correcton algorithm is further optimized -improve image reading source code to accept higher framerates. The hardware capabilities provide image reading frequencies of up to 90fps, current bottleneck is the c++ implementaion which only provides 30 fps max.

-When utilizng higher framerates, the multiproceesng approach will be viable to gain performance. Optimizaton work on the implementaion can be done.

-Image manipulaton on the cpu is rather costly, even at small image sizes. The raspberry Pi has a "small" gpu unit onboard, whch is in dle state when the device is run in headless mode. The gpu could then be utilized to do heavy weight image calculation such as stereo rectification. -Optimization point would also be creating a mor optimized verson of the color thresholding algorithm where the threshodling for all of the thracking colors is done in one run on the current frame. It is also to be evaluated if this problem falls can be expressed as a SIMD function and would benefit from the processng on the GPU.

The current prototype is still missing an graphcal user interface on the display side. Further communication of states on the slave Pi's to the master unit would also be beneficial. The hand data that is used for prototyping is still hard coded and only manually configurable. A calibration procedure for the system as well as a suitable data format for representing and translating these values for the IK algortihm s still to be done.

-Shrnk tubing showed to be a utilzable material for the usecase. Downsinde of the material is that the colors used are the largest spectrum of colors available on the market and geting outer colors is relatively hard. Other material shuold be evaluated.

Furthermore the images are com-pressed with a JPEG algo-rithm which reduces color channel reso-lution. Access to the raw frame data from the camera would be benefi-cial to improve tracker color segmen-tation.

9 Resume

The constructed prototype showed, that it is possible to construct a basic hand and object tracking system with consumer grade hard and software. The limited hardware capabilities of the used Raspberry Pi showed to be able to supply enough processing power to run live image processing operations at up to 30 FPS. A downside of the used camera hardware is that the used sensor utilizes a rolling instead of a global shutter. This makes the synchronization of the two cameras for the stereoscopic setup harder. Furthermore, the cameras were not intended to be used in such a scenario, therefore they lack a frame synchronization feature. This downside has to be corrected on the software side. timimg measurements showed that at lower framerates, the described multithreaded processing approach does not reach much higher processing times than the sequential approach. The sequential process is also less error prone since it does not need the additional synchronization wrappings that the multithreaded approach needs. Should higher readout frame rates be achieved from the camera, the multithreaded approach should definitely be used.

List of Figures

2.1	Bone structure of the human left hand ([30])	2
2.2	Types of degrees of freedom for the joints of the human hand	3
2.3	Possible solution for an IK problem of a human finger:(a)The given target position of the end effector can not be reached. (b) The given target can only be reached by one solution.(c) The target position can be reached with multiple different solutions.	7
2.4	Forward and backward calculation steps for one iteration of the FAB-RIK algorithm.(a) initial position of the system.(b)End effector is moved to target position.(c) Deterimne position of next joint on constructed line.(d)repeat until root is reached. (e)move root joint to initial position. (f) repeat calculation in reverse direction [2]	12
3.1	Example for an optical tracking system with passive infrared reflector markers [41]	15
3.2	Acceleglove system with the described sensor positions [19]	17
3.3	Color glove setup used by Wang and Popovic [53]	19
3.4	Color glove setup used by Fredriksson and Ryen [15]	20
3.5	Microsoft kinect camera system with the three used sensor units[58] . . .	22
3.6	Visualization of a (a) Real (using Infrared Imaging) and (b) Schematic View of Leap Motion Controller[54].	22
3.7	Visualization of a <i>Intel Reas/Sense</i> depth module[22]	23
4.1	Pavlovic, Sharma et al: Hand models. Different hand models can be used to represent the same hand posture. (a) 3D Textured volumetric model. (b) 3D wire frame volumetric model. (c) 3D skeletal model. (d) Binary silhouette. (e) Contour.	24
6.1	Technical conception of the single hardware parts	31
6.2	Technical drawing of the camera holder	33
6.3	Printed camera holder with mounted cameras	34
6.4	Image point projection for pinhole camera systems[18]	34
6.5	ROI calculation for consecutive frames	36
6.6	Example Mask of 3x3 size : a) empty mask b) Mask that keeps pixel value c) combination of b) with a left shift d) combination of b) with an upshift e) isotropic mask	39

List of Figures

6.7	Workflow for the UDP Data packages that are send from the Raspberrys to the parent PC	42
6.8	Overview diagram of the processing steps for the hand model data on the client side	43
6.9	Stereoscopic areas	45
6.10	Dimmensions for calculation.Images from:[24]	46
7.1	Work flow of the C++ implementation with thread pool solution and ROI	51
7.2	Images showing reflection problems with orange color thresholding. a) Correct threshold with only the marker being thresholded b-c) Depending on the rotaton of the hand, skin reflecton color falls into tracking threshold.	53
7.3	First set of shrink tube color markers	54
7.4	Improved larger and smaller set of shrink tube color markers	54
7.5	Different types of tested markers. Thumb with complete acrylic coloring, index and ring finger with smaller shrink tube markers, middle finger with nail polish style acrylic color and little finger with large improved marker .	56
7.6	Chart showing the results of the disparity measurement with a plotted trend line	57
7.7	Hand model representatio of the used IK Framework: a) Kinematic chains and base structure, b) hand model with visualized movement constraints	58

List of Tables

7.1 measurement results for sequential processing timing measurement	52
10.1 Depth measurement Values	xii

Bibliography

- [1] ANDREAS ARISTIDOU AND JOAN LASENBY: Inverse Kinematics: a review of existing techniques and introduction of a new fast iterative solver: University of Cambridge, Department of Engineering. (2009)
- [2] ARISTIDOU, Andreas ; LASENBY, Joan: FABRIK: A fast, iterative solver for the Inverse Kinematics problem. In: *Graphical Models* 73 (2011), Nr. 5, S. 243–260. – ISSN 15240703
- [3] BADLER, Norman ; MANOOCHEHRI, Kamran ; WALTERS, Graham: Articulated Figure Positioning by Multiple Constraints. In: *IEEE Computer Graphics and Applications* 7 (1987), Nr. 6, S. 28–38. – ISSN 02721716
- [4] BOLT, Richard A.: Put-that-there. In: THOMAS, James J. (Hrsg.): *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. New York, NY : ACM, 1980, S. 262–270. – ISBN 0897910214
- [5] CHIAVERINI, S. ; SICILIANO, B. ; EGELAND, O.: Review of the damped least-squares inverse kinematics with experiments on an industrial robot manipulator. In: *IEEE Transactions on Control Systems Technology* 2 (1994), Nr. 2, S. 123–134. – ISSN 10636536
- [6] DAHMEN, Wolfgang ; REUSKEN, Arnold: *Numerik für Ingenieure und Naturwissenschaftler*. Zweite, korrigierte Auflage. Berlin Heidelberg : Springer-Verlag Berlin Heidelberg, 2008 (Springer-Lehrbuch). – URL <http://dx.doi.org/10.1007/978-3-540-76493-9>. – ISBN 9783540764939
- [7] DAVID E. ORIN ; WILLIAM W. SCHRADER: Efficient Computation of the Jacobian for Robot Manipulators. In: *The International Journal of Robotics Research* 3 (1984), Nr. 4, S. 66–75. – ISSN 0278-3649
- [8] DAVIES, E. R.: *Computer Vision: Principles, Algorithms, Applications, Learning*. Academic Press, 2017
- [9] DENAVIT, J. ; R.S.HARTENBERG: A kinematic notation for lower-pair mechanisms based on matrices. In: *ASME J. Appl. Mech.* (1955), S. 215–221
- [10] DIPIETRO, L. ; SABATINI, A. M. ; DARIO, P.: A Survey of Glove-Based Systems and Their Applications. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38 (2008), Nr. 4, S. 461–482. – ISSN 1094-6977

Bibliography

- [11] DUCA, Florin ; FREDRIKSSON, Jonas ; FJELD, Morten: Real-time 3d hand interaction: Single webcam low-cost approach. In: *Proceedings of the Workshop at the IEEE Virtual Reality 2007 Conference; Trends and Issues in Tracking for Virtual Environments*, 2007, S. 1–5
- [12] EROL, Ali ; BEBIS, George ; NICOLESCU, Mircea ; BOYLE, Richard D. ; TWOMBLY, Xander: Vision-based hand pose estimation: A review. In: *Computer vision and image understanding* 108 (2007), Nr. 1-2, S. 52–73
- [13] FOXLIN, Eric ; HARRINGTON, Michael ; PFEIFER, George: Constellation. In: CUNNINGHAM, Steve (Hrsg.): *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. New York, NY : ACM, 1998, S. 371–378. – ISBN 0897919998
- [14] FRANZiska MUELLER ; DUSHYANT MEHTA ; OLEKSANDR SOTNYCHENKO ; SRINATH SRIDHAR ; DAN CASAS ; CHRISTIAN THEOBALT: Real-Time Hand Tracking Under Occlusion From an Egocentric RGB-D Sensor. (2017). – URL http://openaccess.thecvf.com/content_ICCV_2017_workshops/papers/w19/Mueller_Real-Time_Hand_Tracking_ICCV_2017_paper.pdf. – Zugriffsdatum: 20.05.2018
- [15] FREDRIKSSON, Jonas ; RYEN, Sven B. ; FJELD, Morten: Real-time 3D hand-computer interaction. In: TOLLMAR, Konrad (Hrsg.): *Proceedings of the 5th Nordic conference on Human-computer interaction building bridges*. New York, NY : ACM, 2008, S. 133. – ISBN 9781595937049
- [16] GOLUB, Gene ; KAHAN, William: Calculating the singular values and pseudo-inverse of a matrix. In: *ISAM* 2 (1965), Nr. 2, S. 205–224. – URL <http://statistics.uchicago.edu/~lekheng/courses/302/classics/golub-kahan.pdf>. – Zugriffsdatum: 06.10.2017
- [17] GRIMES, Gary J.: *Digital data entry glove interface device*. 1983
- [18] HARTLEY, Richard ; ZISSEMAN, Andrew: Multiple View Geometry in Computer Vision Second Edition. In: *Cambridge University Press* (2000)
- [19] HERNANDEZ-REBOLLAR, Jose L. ; KYRIAKOPOULOS, Nicholas ; LINDEMAN, Robert W.: The AcceleGlove. In: APPOLLONI, Tom (Hrsg.): *ACM SIGGRAPH 2002 conference abstracts and applications*. New York, NY : ACM, 2002, S. 259. – ISBN 1581135254
- [20] HTC: *HTC Vive*. 2018. – URL <https://www.vive.com/de/product/>. – Zugriffsdatum: 10.05.2018

Bibliography

- [21] HUTTENLOCHER, D. P. ; KLANDERMAN, G. A. ; RUCKLIDGE, W. J.: Comparing images using the Hausdorff distance. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15 (1993), Nr. 9, S. 850–863. – ISSN 01628828
- [22] INTEL CORPORATION: *Inter RealSense Technology*. 2018
- [23] JAMIE SHOTTON (Hrsg.) ; ANDREW FITZGIBBON (Hrsg.) ; MAT COOK (Hrsg.) ; TOBY SHARP (Hrsg.) ; MARK FINOCCHIO (Hrsg.) ; RICHARD MOORE (Hrsg.) ; ALEX KIPMAN (Hrsg.) ; ANDREW BLAKE (Hrsg.): *Real-Time Human Pose Recognition in Parts from Single Depth Images*. 2011
- [24] JERNEJ MROVLJE ; DAMIR VRANCIC: Distance measuring based on stereoscopic pictures. In: *9th International PhD Workshop on Systems and Control: Young Generation Viewpoint* (2008). – URL <http://dsc.ijs.si/files/papers/s101%20mrovlje.pdf>. – Zugriffsdatum: 15.04.2018
- [25] KOREIN, James U.: *A geometric investigation of reach*: Zugl.: Pennsylvania Univ., Diss. : 1984. Cambridge, Mass. : MIT Press, 1985 (ACM distinguished dissertations). – ISBN 0262111047
- [26] KRISTENSEN, Fredrik ; NILSSON, Peter ; ÖWALL, Viktor: Background Segmentation Beyond RGB. In: NARAYANAN, P. J. (Hrsg.) ; NAYAR, Shree K. (Hrsg.) ; SHUM, Heung-Yeung (Hrsg.): *Computer Vision - ACCV 2006*. Guildford : Springer London, Feb. 2007, S. 602–612. – ISBN 978-3-540-32432-4
- [27] KURODA, T. ; TABATA, Y. ; GOTO, A. ; IKUTA, H. ; MURAKAMI, M. u.a.: Consumer price data-glove for sign language recognition. In: *Proc. of 5th Intl Conf. Disability, Virtual Reality Assoc. Tech., Oxford, UK*, 2004, S. 253–258
- [28] LA GORCE, Martin de ; PARAGIOS, Nikos ; FLEET, David J.: Model-based hand tracking with texture, shading and self-occlusions. In: *IEEE Conference on Computer Vision and Pattern Recognition, 2008*. Piscataway, NJ : IEEE Service Center, 2008, S. 1–8. – ISBN 978-1-4244-2242-5
- [29] LANSLEY, Alastair ; VAMPLEW, Peter ; SMITH, Philip ; FOALE, Cameron: Caliko: An Inverse Kinematics Software Library Implementation of the FAB-RIK Algorithm. In: *Journal of Open Research Software* 4 (2016), Nr. 1. – URL <http://openresearchsoftware.metajnl.com/articles/10.5334/jors.116/galley/192/download/>. – ISSN 2049-9647
- [30] LEE, Jintae ; KUNII, T. L.: Model-based analysis of hand posture. In: *IEEE Computer Graphics and Applications* 15 (1995), Nr. 5, S. 77–86. – ISSN 02721716

Bibliography

- [31] LIN, John ; WU, Ying ; HUANG, T. S.: Modeling the constraints of human hand motion. In: *Proceedings, Workshop on Human Motion*. Los Alamitos, Calif : IEEE Computer Society, 2000, S. 121–126. – ISBN 0-7695-0939-8
- [32] MAJEAU, L. ; BORDUAS, J. ; LORANGER, S. ; EL-IRAKI, Y. ; LAVOIE, J. ; BANVILLE, D. ; LATENDRESSE, V. ; BELAND, V. ; DANIEL-RIVEST, J. ; THIAW, A. ; BAMBARA, H. ; BEAUSOLEIL, T. P. ; TROTTIER-LAPOINTE, W. ; LAPPOINTE, J.: Dataglove for consumer applications: Low cost dataglove using optical fiber sensor. In: *7th International Workshop on Fibre and Optical Passive Components (WFOPC), 2011*. Piscataway, NJ : IEEE, 2012, S. 1–4. – ISBN 978-1-4577-0861-9
- [33] MANAF A. MAHAMMED ; AMERA I. MELHUM ; FARIS A. KOCHERY: Object Distance Measurement by Stereo VISION. In: *International Journal of Science and Applied Information Technology (IJSAIT)* 2 (2013)
- [34] MOESLUND, Thomas B. ; GRANUM, Erik: A Survey of Computer Vision-Based Human Motion Capture. In: *Computer vision and image understanding* 81 (2001), Nr. 3, S. 231–268
- [35] MOESLUND, Thomas B. ; HILTON, Adrian ; KRÜGER, Volker: A survey of advances in vision-based human motion capture and analysis. In: *Computer vision and image understanding* 104 (2006), Nr. 2-3, S. 90–126
- [36] NAKAMURA, Yoshihiko ; HANAFUSA, Hideo: Inverse Kinematic Solutions With Singularity Robustness for Robot Manipulator Control. In: *Journal of Dynamic Systems, Measurement, and Control* 108 (1986), Nr. 3, S. 163. – ISSN 00220434
- [37] OCULUS VR, LLC.: *occulus rift*. 2018. – URL <http://www.oculusvr.com/rift>. – Zugriffsdatum: 10.05.2018
- [38] OIKONOMIDIS, Iason ; KYRIAZIS, Nikolaos ; ARGYROS, Antonis: Efficient model-based 3D tracking of hand articulations using Kinect. In: HOEY, Jesse (Hrsg.): *Proceedings of the British Machine Vision Conference 2011*. Durham : BMVA Press, 2011, S. 101.1–101.11. – ISBN 1-901725-43-X
- [39] OPENCV: *Camera Calibration*. 2018. – URL https://docs.opencv.org/3.1.0/dc/dbb/tutorial_py_calibration.html. – Zugriffsdatum: 10.02.2018
- [40] OPENCV TEAM: *OpenCV*. 2018. – URL <https://opencv.org/>. – Zugriffsdatum: 10.05.2018
- [41] OPTITRACK: *flex13MocapVolume.png (680×375)*. 2017. – URL <http://www.optitrack.com/public/images/flex13MocapVolume.png>. – Zugriffsdatum: 29.09.2017

Bibliography

- [42] ORIN, David E. ; SCHRADER, William W.: Efficient Computation of the Jacobian for Robot Manipulators. In: *The International Journal of Robotics Research* Vol.3 (1984), Nr. No.4. – ISSN 0278-3649
- [43] PAVLOVIC, V. I. ; SHARMA, R. ; HUANG, T. S.: Visual interpretation of hand gestures for human-computer interaction: A review. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (1997), Nr. 7, S. 677–695. – ISSN 01628828
- [44] RAAB, Frederick ; BLOOD, Ernest ; STEINER, Terry ; JONES, Herbert: Magnetic Position and Orientation Tracking System. In: *IEEE Transactions on Aerospace and Electronic Systems* AES-15 (1979), Nr. 5, S. 709–718. – ISSN 0018-9251
- [45] RIJPKEMA, Hans ; GIRARD, Michael: Computer animation of knowledge-based human grasping. In: THOMAS, James J. (Hrsg.): *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*. New York, NY : ACM, 1991, S. 339–348. – ISBN 0897914368
- [46] ROLLAND, Jannick P. ; BAILLOT, Yohan ; GOON, Alexei A.: A SURVEY OF TRACKING TECHNOLOGY FOR VIRTUAL ENVIRONMENTS: Center for Research and Education in Optics and Lasers (CREOL) University of Central Florida, Orlando FL 32816. In: *(Fundamentals of wearable computers and augmented reality* (2001)
- [47] SPONG, Mark W. ; VIDYASAGAR, Mathukumalli: *Robot dynamics and control*. John Wiley & Sons, 2008
- [48] STAUFFER, C. ; GRIMSON, W.E.L.: Adaptive background mixture models for real-time tracking. In: *1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos, Ca : IEEE Computer Society, 1999, S. 246–252. – ISBN 0-7695-0149-4
- [49] STURMAN, D. J. ; ZELTZER, D.: A survey of glove-based input. In: *IEEE Computer Graphics and Applications* 14 (1994), Nr. 1, S. 30–39. – ISSN 02721716
- [50] TAUER, Holger: *Stereo-3D: Grundlagen, Technik und Bildgestaltung*. Fachverlag Schiele & Schoen, 2010
- [51] THOMAS A. DEFANTI ; DANIEL J. SANDIN: *Final Report to the National Endowment of the Arts*
- [52] WAMPLER, Charles: Manipulator Inverse Kinematic Solutions Based on Vector Formulations and Damped Least-Squares Methods. In: *IEEE Transactions on Systems, Man, and Cybernetics* 16 (1986), Nr. 1, S. 93–101. – ISSN 0018-9472

Bibliography

- [53] WANG, Robert Y. ; POPOVIĆ, Jovan: Real-time hand-tracking with a color glove. In: *ACM transactions on graphics (TOG)* Bd. 28, 2009, S. 63
- [54] WEICHERT, Frank ; BACHMANN, Daniel ; RUDAK, Bartholomäus ; FISSELER, Denis: Analysis of the accuracy and robustness of the leap motion controller. In: *Sensors (Basel, Switzerland)* 13 (2013), Nr. 5, S. 6380–6393
- [55] WELCH, G. ; FOXLIN, E.: Motion tracking: No silver bullet, but a respectable arsenal. In: *IEEE Computer Graphics and Applications* 22 (2002), Nr. 6, S. 24–38.
– ISSN 02721716
- [56] WOLOVICH, W. ; ELLIOTT, H.: A computational technique for inverse kinematics. In: *The 23rd IEEE Conference on Decision and Control*, IEEE, 1984, S. 1359–1363
- [57] YASIR DAWOOD SALMAN (Hrsg.) ; KU RUHANA KU-MAHAMUD (Hrsg.) ; EIJI KAMIOKA (Hrsg.): *Distance measurement for self-driving cars using stereo camera in Zulikha*. 2017
- [58] ZHANG, Zhengyou: Microsoft Kinect Sensor and Its Effect. In: *IEEE Multimedia* 19 (2012), Nr. 2, S. 4–10

Eidesstattliche Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben.
Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben.
Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Köln, xx. August 2016

Oliver Kalbfleisch

10 Appendix

Table 10.1: Depth measurement Values

Real distance from camera (cm)	Left camera x-value	Right camera x-value	Disparity in pixels	Calculated distance in cm	Deviation in %(abs.)	Corrected Values	Deviation in %(abs.)
100	302	342	40	99.463	0.54	99.670	0.33
95	301	343	42	94.727	0.29	94.760	0.25
90	300	344	44	90.421	0.47	90.304	0.34
85	298	346	48	82.886	2.49	82.523	-2.91
80	296	345	49	81.194	1.49	80.780	0.98
75	296	348	52	76.510	2.01	75.960	1.28
70	293	349	56	71.045	1.49	70.349	0.5
65	289	350	61	65.222	0.34	64.388	0.94
60	289	355	66	60.281	0.47	59.344	1.09
55	286	357	71	56.036	1.88	55.022	0.04
50	285	361	76	52.349	4.7	51.279	2.56
45	280	365	85	46.806	4.01	45.668	1.48
40	265	362	97	41.016	2.54	39.831	0.42
35	258	367	109	36.5	4.29	35.300	0.86
30	244	373	129	30.841	2.80	29.650	1.17
25	288	381	153	26.003	4.01	24.848	0.61
20	206	395	189	21.050	5.25	19.965	0.17