UNISINOS

Programa de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Rodrigo Bazo

BAPTIZO: A SENSOR FUSION BASED MODEL FOR TRACKING THE IDENTITY OF HUMAN POSES

São Leopoldo, 2019

Rodrigo Bazo

BAPTIZO: A SENSOR FUSION BASED MODEL FOR TRACKING THE IDENTITY OF HUMAN POSES

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Dr. Cristiano André da Costa

São Leopoldo
2019

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)

Bazo, Rodrigo

Baptizo: a sensor fusion based model for tracking the identity of human poses / Rodrigo Bazo — 2019.

71 f.: il.; 30 cm.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2019.

"Advisor: Prof. Dr. Cristiano André da Costa, Unidade Acadêmica de Pesquisa e Pós-Graduação".

1. Sensor Fusion. 2. Tracking. 3. Radio Frequency. 4. Computer Vision. 5. Human Pose Estimation. I. Título.

CDU 004.932

Bibliotecária responsável: Alessandro Dietrich — CRB 10/2338

Rodrigo Bazo

BAPTIZO: A SENSOR FUSION BASED MODEL FOR TRACKING THE IDENTITY OF HUMAN POSES

Dissertação apresentada à Universidade do Vale do I
dos Sinos – Unisinos, como requisito parcial pa
obtenção do título de Mestre em Computação Aplica

Aprovado em 13 de Agosto de 2019

BANCA EXAMINADORA

Gerson Geraldo Homrich Cavalheiro – Universidade Federal de Pelotas

Rodrigo da Rosa Righi – Universidade do Vale do Rio dos Sinos

Prof. Dr. Cristiano André da Costa (Orientador)

Visto e permitida a impressão
São Leopoldo

Prof. Dr. Rodrigo da Rosa Righi
Coordenador PPG em Computação Aplicada

ACKNOWLEDGEMENTS

To Father, to Mother, to Sister;
To Godfathers, To Godmothers, To Fátima, to Izabela P., To Pedro, to Pietro, to Tairini;
To Eduardo F., to Élcio, to Gabriel H., to Rafael K., to Wagner;
To Anselmo, to Brian, to Gabriel S., to Julio, to Mateus, to Rafael P., to Rebecca;
To Eduardo R., to Fernanda, to Lucas, to Rodolfo, to Vinicius;
To Ana Paula, to Eric, to Izabela Q., to Mônica;
To Priscilla, to Ronas, to Tauãnara, to William, to Willian;
A heartfelt thanks.

Ticking away the moments that make up a dull day;
You fritter and waste the hours in an offhand way.
(Pink Floyd – Time)

ABSTRACT

Recent advances in the capabilities of computing devices enable new methods to estimate the pose of humans. Human pose estimation techniques are relevant for several industry fields, such as surveillance and interactive entertainment. Further, encoded human poses provide a valuable input for behavioral analysis and activity recognition. Body part detectors offer millimetric accuracy thanks to state-of-the-art Computer Vision technology. However, they still suffer from issues, such as long-term occlusion, that hinder the identification of human subjects. Such problems are intrinsic to Computer Vision devices and can only be solved either with the use of heuristic methods or the deployment of more cameras, which are not always feasible. In turn, radiofrequency-based tracking systems do not suffer from occlusion or identity loss problems and, albeit not as precise as Computer Vision methods, can achieve a high accuracy level. Radiofrequency positioning systems and human pose estimation techniques can complement each other in different ways. For example, the prior can help to identify tracked humans and reduce occlusion errors while the later can increase the accuracy of obtained positions. Thus, the combination of radiofrequency-based positioning and computer vision-based human pose estimation yields a solution that provides better tracking results. Therefore, this thesis proposes a system that generates identified pose data by fusing the unique identities of radiofrequency sensors with unidentified body poses while using estimated body parts for reducing radiofrequency position estimations errors. Experiments with a proof-of-concept demonstrate the feasibility of the sensor fusion technique. Furthermore, experiments analyzing the proposed error reductiong strategy conducted in a experimentation laboratory and a real operating room also show a potential reduction on positioning errors by nearly 46%.

Keywords: Sensor Fusion. Tracking. Radio Frequency. Computer Vision. Human Pose Estimation.

RESUMO

Os recentes avanços no poder computacional de dispositivos permitem a utilização de novos métodos para a estimativa de poses humanas. Tais técnicas são relevantes para diversos setores da indústria, como segurança e entretenimento. Além disso, poses humanas são um input valioso para análise comportamental e reconhecimento de atividades. Reconhecedores de partes de corpo humana, utilizados em estimativas de pose humana, possuem precisão milimétrica devido aos equipamentos de estado da arte de visão computacional. Porém, estes equipamentos possuem limitações como a oclusão, que dificulta a identificação de pessoas. Tais problemas são nativos aos dispositivos de visão computacional devido a sua natureza, e somente podem ser superados utilizando heuristicas ou aumentando o numero de câmeras, o que não é sempre viável. Por outro lado, sistemas de rastreamento baseados em radiofrequência não sofrem com oclusão ou problemas como perda de identidade, e também alcançam altos níveis de precisão mesmo não sendo tão precisos quanto métodos de visão computacional. Sistemas de rastreamento baseados em radiofrequência e estimativas de pose humanas podem se complementar de diversas maneirars. Por exemplo, o primeiro pode ajudar na identificação de poses estimadas, e as poses podem ser utilizadas para mitigar os erros obtidos. Desta maneira, a combinação de ambas as tecnologias oferecem um resultado de rastreamento de poses com precisão superior. Esta dissertação propõem um sistema que gera poses identificadas, baseado na fusão de identificadores de radiofrequência com poses obtidas através de técnicas de visão computacional. Além disso, uma técnica para redução de erro na estimativa da posição dos dispositivos de radiofrequência utilizando poses estimadas é proposta. Experimentos demonstram a viabilidade da fusão de ambos tipos de dados. Além disso, reduções de erros de até 46% utilizando a estratégia de redução de erro proposta são observados. Tanto em experimentos conduzidos em um laboratório de experimentação quanto em uma sala cirúrgica real.

Palavras-chave: *Sensor Fusion*. Rastreamento. Radiofrequência. Visão Computacional. Estimativa de Pose Humana.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AoA	Angle of Arrival
API	Application Programming Interface
BLE	Bluetooth Low-Energy
BSN	Body Sensor Network
CNN	Convolutional Neural Network
CV	Computer Vision
FoV	Field-of-View
FPS	Frames Per Second
GPS	Global Position System
HPE	Human Pose Estimation
IoT	Internet of Things
NFC	Near Field Communication
OR	Operating Room
OS	Operating Suite
RF	Radiofrequency
RFID	Radiofrequency Identification
RGBD	RGB Depth
RMSE	Root Mean Square Error
RSS	Received Signal Strength
RTLS	Real-Time Location System
SL	Structured Light
SV	Stereo Vision
TDoA	Time-Difference of Arrival
ToA	Time of Arrival
ToF	Time of Flight
UWB	Ultrawide-band
WSN	Wireless Sensor Network

CONTENTS

1 INTRODUCTION

Detection and categorization of human activities is a continuously growing research field of Computer Science. Such techniques aim to automatize human activity recognition using data collected from different types of sensors. Usually, solutions are developed to recognize a small set of predefined human activities, which directly impact the selection of employed sensing technologies. In turn, more sophisticated activities require a richer dataset provided by more sensors. Recent advances on the capabilities of computing devices and the consequent evolution of Convolutional Neural Networks (CNN) enable new methods to estimate the pose of humans (ANTUNES et al., 2018). Human Pose Estimation (HPE) is a technique of great interest to several industry sectors such as surveillance, digital entertainment, driving assistance (XIU et al., 2018). Furthermore, cost-efficient RGB and RGB-Depth (RGBD) devices also became available, further stimulating research on the HPE field (AGGARWAL; XIA, 2014).

However, HPE techniques suffer from various forms of occlusion. Moving objects have a high chance of blocking the Field-of-View (FoV) of cameras, leading to occlusion and loss of identity problems (IQBAL; MILAN; GALL, 2017). Such issues can only be solved by either deploying more cameras or implementing heuristics such as temporal consistency. These solutions are not always feasible due to the high cost and intrusiveness involved. However, human activities are highly contextual, therefore issues related to identification are a key challenge related to activity recognition.

Indoor radiofrequency (RF) based tracking solutions, also known as Indoor Positioning Systems (IPS) or Real-Time Location Systems (RTLS) (BOULOS; BERRY, 2012; HAUTE et al., 2016), are an alternative solution of identity and location tracking. Many RTLS available on the market enable identification and position tracking with sub-room accuracy levels. These systems do not suffer from occlusion and can monitor a larger area compared to RGBD cameras. In comparison, location systems based on Computer Vision (CV) can achieve a higher precision, albeit limited to the cameras' FoV.

It is possible to combine the data collected from RTLS and HPE systems and, consequently, obtain a location and pose tracking solution with higher accuracy (MANDELJC et al., 2012). The fusion of multi-sensor data from RF and RGBD sensors brings several advantages (GRAVINA et al., 2017), such as: increased confidence, enhanced robustness and improved precision. Each detected human pose can be associated with an identity based on the information collected from RTLS tags. Further, the millimiter accuracy of state-of-the-art RGBD cameras (YANG et al., 2015; LACHAT et al., 2015) provide highly accurate location information for objects in their FoV. This information can then be used for reducing the estimation error of RF tags position.

1.1 Motivation

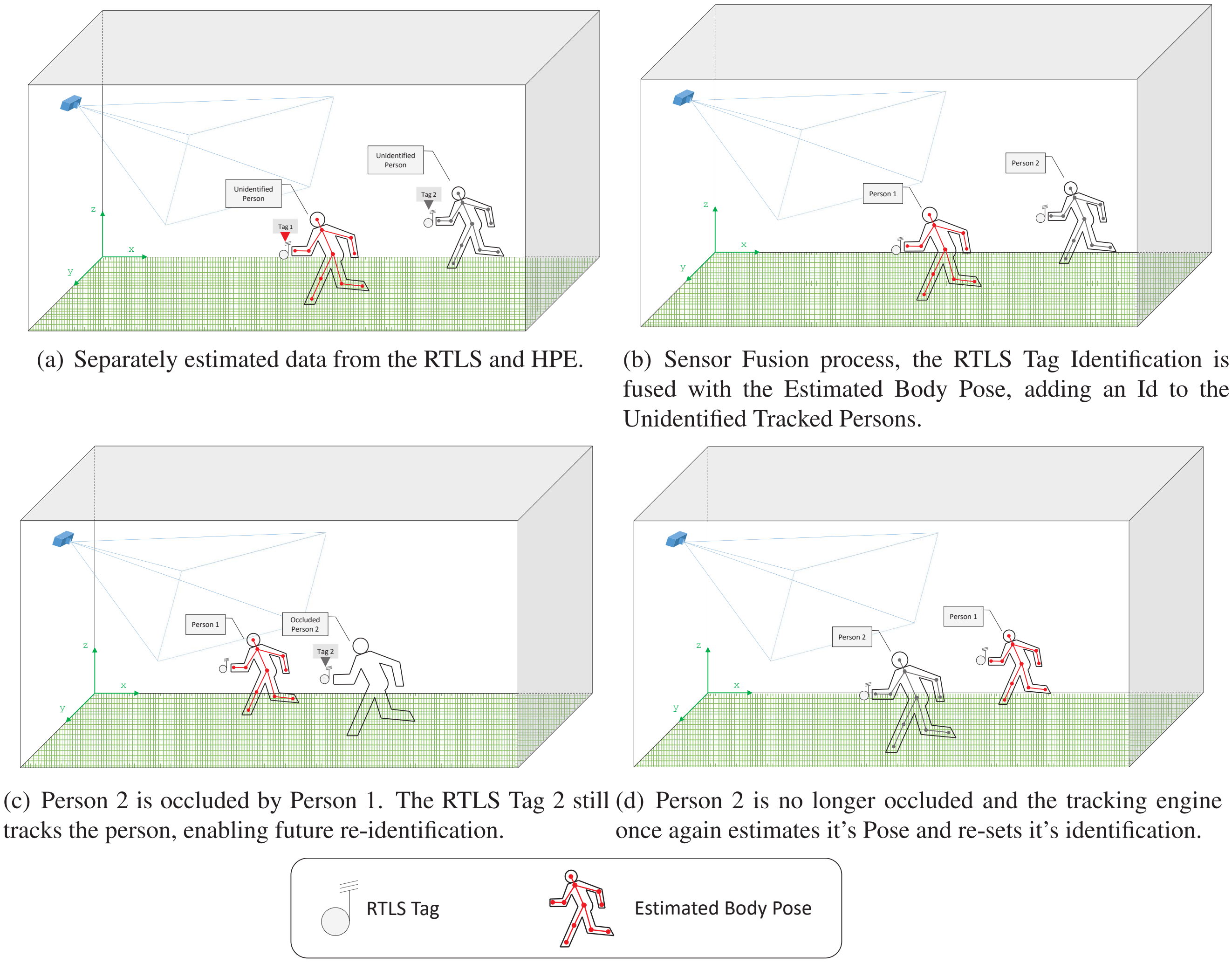
According to (HIGHTOWER; BORRIELLO, 2001), in the context of location systems, Sensor Fusion refers to the technique of obtaining most accurate and precise location data by fusing different location systems data. This kind of correlation can be used for: improving logistics by workflow monitoring (ANTUNES et al., 2018), obtaining valuable data by correlating patient location data and it's EHR (KANG et al., 2014; SHIREHJINI; YASSINE; SHIRMOHAMMADI, 2012), enabling infectious disease control by contact tracing (HELLMICH et al., 2017). Further, (HIGHTOWER; BORRIELLO, 2001) state that, the more independent to each other the location technique used, the more effective the combination is.

While providing highly accurate position data, CV-based tracking applications lack the ability to consistently maintain tracked subjects identity even on marker-based approaches, due to the commonly chaotic scenario of people's movements and occlusions. Radiofrequency-based applications on the other hand, consistently read the target's unique id and estimate it's position if the tag is under the designed coverage area (MANDELJC et al., 2012). Considering the complementary nature of both methods, it is only natural to combine them for obtaining better results. Furthermore, a tendency is rising regarding the fusion of CV and RTLS data (ANTUNES et al., 2018).

Applying Data Fusion on the RTLS and CV-generated data, it is possible to keep track of the identity of tracked humans by fusing these data. Applying a set of heuristics to both obtained data, the precision of both devices can be enhanced. Furthermore, the RTLS allows recovery from inevitable occlusion cenarios that CV-based tracking methods encounter, as illustrated in Figure 1. On a first moment, both RTLS and HPE data are separately estimated, afterwards their data is fused. However, the fused data can can be compromised due to unexpected occlusions and missing pose estimations. The RTLS Tag position and identification enable re-identification of occluded person's poses. This way the data fusion quality can be further improved, generating most reliable and robust information of tracked human poses with identification, enabling applications that take advantage of this data, such as activity recognition, work on top of quality data. Furthermore, this model could be of great use for areas such as robotics, surveillance and augmented or virtual reality applications. Such data correlation can even lead to computational resources managing benefits, like: turning on and enabling processing only of cameras that have RF tags on it's Field of View (FoV) or providing RTLS location data as input for the boosting the pose estimation performance.

1.2 Objectives

In this context, this thesis proposes the Baptizo model. It enables human pose tracking based on the fusion of RTLS location and identificatin data with human poses estimated by a CV-based HPE strategy. This work details the model's architecture, including its modules,



(a) Separately estimated data from the RTLS and HPE.

(b) Sensor Fusion process, the RTLS Tag Identification is fused with the Estimated Body Pose, adding an Id to the Unidentified Tracked Persons.

(c) Person 2 is occluded by Person 1. The RTLS Tag 2 still tracks the person, enabling future re-identification.

(d) Person 2 is no longer occluded and the tracking engine once again estimates it's Pose and re-sets it's identification.

Figure 1: Fusing the data from the HPE and RTLS enables identifying extracted body poses, while still tracking occluded persons and re-identifying them after they go unnocluded.

components, and heuristics used for fusing sensor data. Results demonstrate that, among other conclusions, the model reduces errors on RTLS tags position estimation by a significant margin. The main contributions of this work include:

- The sensor fusion model to combine RTLS and HPE data, which is not explored in the literature to the current date;
- A novel technique for reducing RF-based position estimation errors through sensor fusion with body parts estimation data.

1.3 Text Structure

The remainder of this thesis is organized as follows: in Chapter 2 a review of the basic aspects of components and techniques of both RF tracking systems and CV for poses estimations is conducted; Chapter 3 conducts a review of relevant state-of-the-art works, discussing and analyzing similar works and trending topics on the areas that compose this thesis; the envisioned Baptizo model is detailed in Chapter 4; the methodology used for testing the proposed prototype is detailed in Chapter 5; Chapter 6 details the experiments results obtained with the proposed methodology in different scenarios with different number of subjects;finally, Chapter 7 concludes this thesis proposal.

2 BACKGROUND

This work aims at fusing the data from two different data sources: RTLS location data and human poses generated via HPE. In this chapter, essential concepts, technologies and techniques regarding the beforementioned technologies will be analyzed and discussed in order to understand challenges and guaranteeing use of the most suited technologies and techniques on the proposed model. In Section 2.1, important aspects of RTLS, such as used devices and strategies for location estimation, will be reviewed. And, in Section 2.2, basic concepts, challenges, techniques and devices used for conducting HPE will be studied, as well as the principles behind RGBD cameras. Lastly, Section 2.3 quickly conducts a review on data fusion.

2.1 Real-Time Location Systems

Ever since 2001, when the notorious work of (HIGHTOWER; BORRIELLO, 2001) was published, the GPS was already the most publicized location-sensing system, and it's popularity has grown non-stop up until today. While location-based services are already ubiquitous on GPS-enabled devices, such as smartphones, they are limited to outdoor environments since GPS functionality is compromised in indoor environments (HIGHTOWER; BORRIELLO, 2001; HAUTE et al., 2016; BOULOS; BERRY, 2012). RTLS, also called IPS, are local systems that enable identifying and tracking assets or personnel in indoor environments. Using the same main principle from the GPS, RTLS consist of a set of anchors and a set of mobile nodes attached or carried by tracked entities. The mobile nodes on RTLS are most usually referred to as tags or sensors. In Figure 2, the principle behind both technologies is illustrated. In contrast with the GPS which estimated the node position using satellites as anchors, the RTLS anchor nodes are devices deployed in indoor spaces that receive the signal emitted by the tags. While RTLS can be implemented using different kind of technologies such as infrared and ultrasound, the majority of commercial or scientific RTLS are based on radio frequency (RF). Considering that, RTLS will be treated as radio-frequency based RTLS on this work. Such location systems can be implemented in many ways with different location techniques and RF devices, and the used techniques and devices depend mostly on the desired accuracy level. Lastly, most RTLS work under the Wireless Sensor Network (WSN) concept. WSN are networks composed of several small devices that intraoperate through a set of protocols in order to collect information about the environement (RAWAT et al., 2014).

2.1.1 RTLS-enabling Sensors

Several technologies enable tracking and identifying assets within indoor spaces, however RF-based technologies are trending in RTLS context. RF devices enable tracking and identifying entities carrying tags in indoor spaces. Within the RF spectrum, several devices exist, and

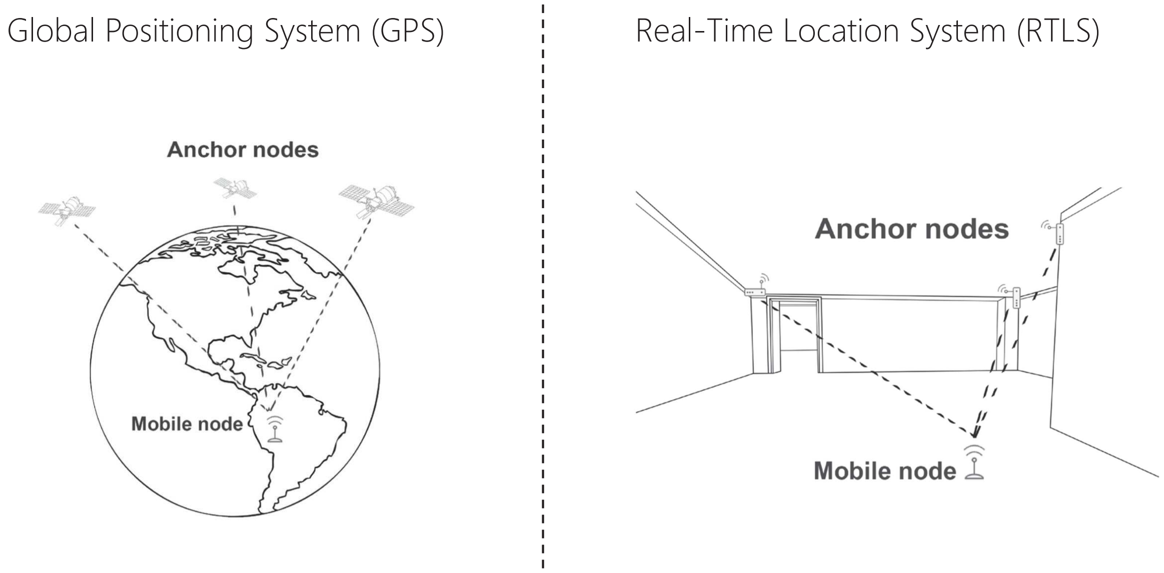

Global Positioning System (GPS)
Real-Time Location System (RTLS)
Anchor nodes
Mobile node
Anchor nodes
Mobile node

Figure 2: Comparison of Global Positioning System and Real-Time Location System. Source: Adapted from (HAUTE et al., 2016).

will be detailed next. RF-based systems usually consist of a tag that emits it's unique id and a set of anchors that receive the tags signal. In Figure 3, the functioning of a RTLS based on RF devices is illustrated. The tag emits it's unique id signal that is collected by the Anchors and consumed by a RTLS Server that applies a localization algorithm in order to estimate the tag location.

Radio Frequency Identification (RFID) is probably the most well known RF technology and is categorized as Active RFID or Passive RFID. Passive RFID, constantly referred to simply as RFID and so on this text, is a consolidated RF device that aims at providing identification in short-range, such solution is extensively explored in the literature. RFID tags have very low cost and require no energy source, since in this case, the anchor's emits the signal that energizes the tag. While having many desirable characterstics, it is not a suitable technology for employing on a high accuracy RTLS, since it's range is short, the tag must be close to the reader. Further, it's localization principle relies on proximity-based algorithms for estimating the tag position, which has poor accuracy and are mostly used for identifying entities in small range(BOULOS; BERRY, 2012; HAUTE et al., 2016).

On the other hand, Active RFID is an expensive technology and it's tags require a power source. Active RFID tags constantly emits it's signal and the singal range is much greater than it's Passive counterpart, reaching up to 100 meters (ZHAO; LIU; NI, 2007). Still inside the RFID spectrum, Near Field Communication (NFC) is a RFID based technology that works in a similar fashion to RFID, only working in different frequencies and with shorter signal range. While NFC is a broadly used technology in various mobile systems (LAHTELA; HASSINEN; JYLHA, 2008), it isn't a suitable technology for high accuracy-demanding RTLS as needed in this thesis's proposed model.

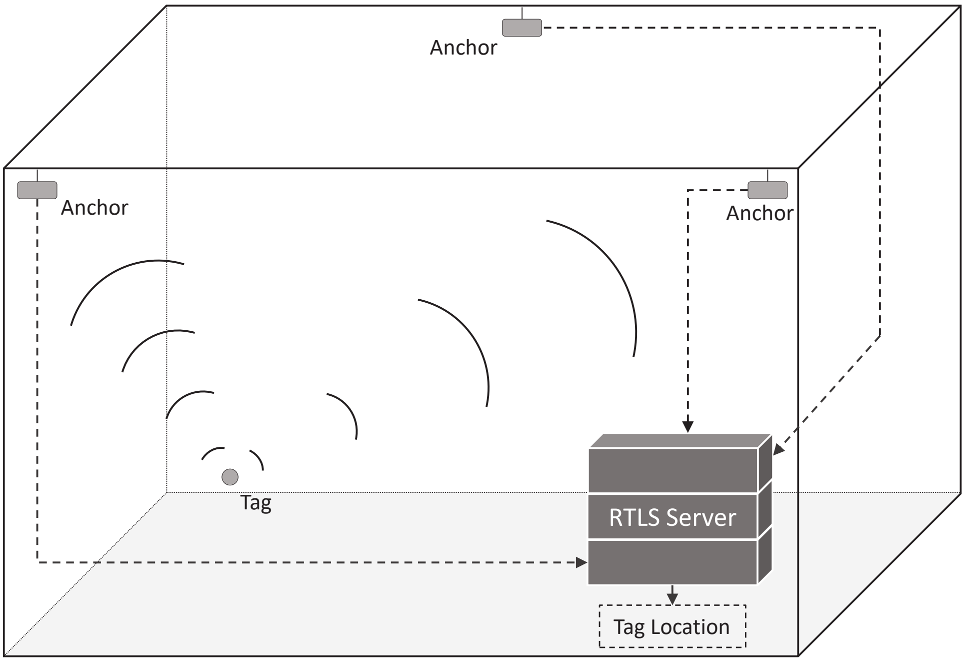


Figure 3: Real-Time Location System basic architecture.

Another RF technology is the Ultra-wideband (UWB). The UWB cyclically transmits information in very short pulses. In relation to RFID, UWB operates on multiple bands of frequencies simultaniously, floating from 3.1 to 10.6 GHz. Further, it is a low energy consuming technology and robust towards interference with other RF devices. Regarding location estimation, the UWB is particularly interesting. Due to it's principle of short duration pulses, filtering correct and multipath generated signals is made easier (LIU et al., 2007). Despite it's high costs and unsuitability for long range communications, the UWB is a promising technology for high accuracy RTLS and WSN (RAWAT et al., 2014).

Together with the UWB, Bluetooth and WiFi are part of the short-range wireless field (RAWAT et al., 2014). WiFi is well known and highly present protocol that allows high range data transmission with large data throughput, however requiring high energy consumption, while Bluetooth focuses on short range wireless communication on low-cost devices. On version 4.0, the Bluetooth Special Interest Group standardized Bluetooth Low-Energy (BLE). Which is an interesting choice for WSN applications that require high data transmission in short distances between devices. Finally, similarly, ZigBee is a wireless communication technology usefull for applications that focus on low energy consumption and cost (ANTUNES et al., 2018). In Table 1, technical details of the described communication technologies are presented.

2.1.2 Localization Techniques

Typically, RTLS run on the top of an algorithm that processes data from a specific technology. Therefore, RTLS can be seen as a combination of a given localization techniques and a

Table 1: Described communication technologies technical details.

Technology	RFID	NFC	WiFi	Bluetooth	UWB	ZigBee	BLE
Specification	ISO 15693, ISO 14443, ISO 18000	ISO 14443, ISO 18092	IEEE 802.11	IEEE 802.15.1	IEEE 802.15.3a	IEEE 802.15.4	IEEE 802.15.4
Frequency band	< 100 MHz, 868 MHz, 915 MHz, 2.45 GHz	13.56 MHz	2.4 GHz, 5 GHz	2.4 GHz	3.1-10.6 GHz	868-928 MHz, 2.4 GHz	2.4 GHz
Max signal rate	-	424 Kbps	54 Mbps, 540 Mbps	3 Mbps	110 Mbps	250 Kbps	1Mbps
Nominal range	30 cm, 1 m, 3-5 m	10 cm	100 m	10-100 m	10 m	10-100 m	200 m

Source: Adapted from (ANTUNES et al., 2018).

given RF hardware technology. However, not only local position tracking systems such as RTLS suffer from measurement noise, GPS as well. Therefore filtering of raw measurements turn out to be necessary (KAUTZ; GROH; ESKOFIER, 2016). In this subsection, both Localization Techniques and Filters will be explained.

In Figure 4, a classification of location detection is presented according to FARID; NORDIN; ISMAIL (2013) and HAUTE et al. (2016). Next, these classification and techniques will be detailed. One of the simplest localization methods to implement, *Proximity Detection* locates mobile nodes using the highest Received Signal Strength (RSS). This approach is constantly used on applications that use RFID or NFC sensors due to the short range nature of these sensors. However, the implementation complexity of proximity techniques is inversely proportional to it's accuracy, and not even room level accuracy can be obtained with it's application.

Triangulation techniques are based on the geometrical properties of triangles in order to estimate the nodes locations. It is divided into two further categories, *Direction Based* and *Distance Based* techniques. This categories are divided in that fashion considering that algorithms under each category apply their heuristics based on lateration and angulation. Techniques based on RSS or time-propagation measurements fall under the lateration category, also considered *Distance Based* strategies. Algorithms based on the measurement of received signal's angles are on the angulation category, and are listed under *Angle Based*.

Under *Angle Based* techniques the most prominent example is the Angle of Arrival (AoA). This algorithm, as the name suggests, determines mobile nodes signals angle of arrival on the anchors. However, the largest the required accuracy on applications that apply AoA, the more anchors are needed, this way increasing the application cost. Further, in the context of RTLS, this technique is affected by multipath and non-line of sight propagation due to uncertainties of indoor environments. Such characteristics are harmful on AoA position estimation.

On *Distance Based* Techniques, techniques are divided under time or signal based. The most well known time based techniques are Time of Arrival (ToA) and Time Difference of Arrival

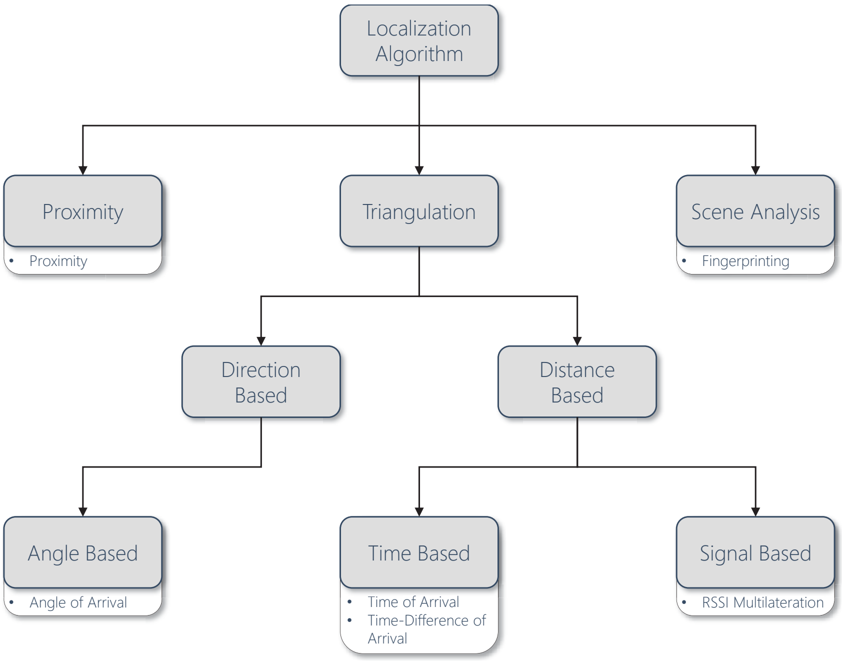
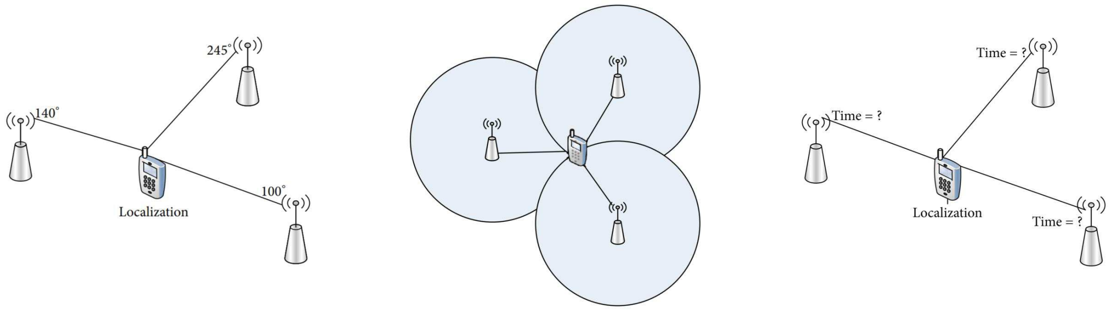


Figure 4: Classification of localization techniques together with most accurate algorithms. Gray boxes are classifications and white boxes represent algorithms under the respective category. Source: Adapted from (FARID; NORDIN; ISMAIL, 2013; HAUTE et al., 2016).

(TDoA). ToA, or also Time of Flight (ToF) uses the signal speed as the main parameter for estimating the mobile node location. When the signal is received by the anchors, the distance from the tag to the anchors is calculated by measuring the time of the radio signal to travel from the tag to the anchors and from the anchors to the tag. While largely accurate, this method requires that the tags not only be able to propagate their signal, but also receive signals. Further, the tag and anchors must be accurately synchronized. On the other hand, the TDoA technique calculates the distance from the tag by calculating the difference in the tag signal arriving time on the anchors. While still very accurate, applications that employ TDoA do not require their tags to receive signals, only to propagate. Also, only the anchors must be kept working in sync. This characteristics make TDoA one of the most suitable localization technique for RTLS that require point level accuracy. On Figure 5, the AoA, ToA and TDoA localization techniques are presented. Each image shows the data used and used by the anchors on the position estimation.

Lastly, *Scene Analysis* algorithms consist of mapping the real world into a data persistency and, afterwards, obtained read wireless data from mobile nodes is compared with the mapped data from the persistency. While highly accuracte, these kind of methods are impracticable for RTLS, which are to be employed in highly convluted and chaotic nature. These techniques hard calibration, demanding that each change on the real world to be manually rerecorded into the persistancy. The most typical example of Scene Analysis localization technique is Fingerprinting.

Drifting the focus from localization techniques to filters, Kalman Filter (KF) is a widely adopted filtering solution that is applied in several ways. KF provides means for real-time fil-



(a) Angle of Arrival (b) Time of Arrival (c) Time-Difference of Arrival

Figure 5: Localization techniques illustrations: Angle of Arrival (a) uses the angle of arrival of the sensor signal as parameter for estimating the tag location. Time of Arrival (b) estimates the tag position based on the time the tag waves take forth and back from the tag to the anchors. Time-Difference of Arrival (c) uses the difference in the time of arrival of the tag signal on the anchors to estimate it's position. Source: Adapted from (FARID; NORDIN; ISMAIL, 2013).

tering, it is a lightweight filter in both memory and processing complexities. The KF algorithm works as follows: for each point k, a *prediction* and a *update* step are executed in order to understand the state x of the analyzed system. The *prediction* step consists of calculating a previous state of the system x, therefore for the a point k, calculating $x_{k-1|k}$. When point $k+1$ reading becomes available, it's next estimated state $x_{k|k+1}$ is estimated according to the last calculated $x_{k-1|k}$ state. Lastly, the update step is executed. While the KF provides a near optimal solution for filtering noises out of non-linear systems, several extensions of it exist such as the Extendend Kalman Filter (EKF), Unscented Kalman Filter (UKF) and the Square-Root Unscented Kalman Filter (SRUKF). While being the most used, EKF suffers from some linearization errors, which can be avoided by using UKF. Lastly, SRUKF was proposed in order to fix some numerical instability that are innate to the UKF solution (KAUTZ; GROH; ESKOFIER, 2016).

2.2 Computer Vision

With the huge increase of computer-processing capabilities in the recent years, CV-based techniques advanced at a much larger pace than other research areas. These capabilities enable the application of several algorithms that were prohibitive complexity-wise. Such methods enable analyzing images and extracting a rich amount of information at real-time. CV is huge area with even larger amount of content, therefore the literature review conducted in this section will have it's scope limited to areas of interest. In the extension of this work, CV techniques for estimating human poses as well as an analysis on RGBD cameras are of great interest, therefore

.

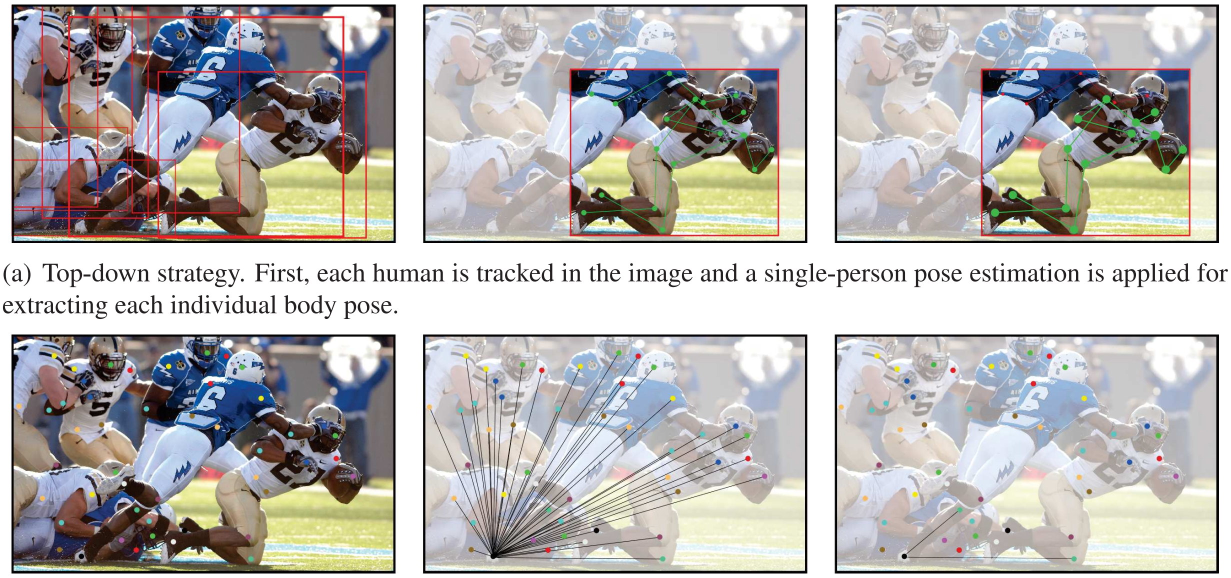
2.2.1 Human Pose Estimation

Estimating human poses is a fundamental task for many trending Computer Vision applications. While many milestones have been achieved recently with the increase in processors capabilities, several challenges, such as occlusion and unexpected poses are still troublesome (GIRSHICK et al., 2011). A well established way to overcome such challenges is the use of markers in order to keep track of occluded and the identity of tracked subjects (HOLTE et al., 2012). However, marker-based solutions are intrusive and undeployable in unconstrained scenarios. Markerless HPE can be defined as determining precise pixel location of body keypoints, followed by the estimation of the human subjects' poses (NEWELL; YANG; DENG, 2016). Such technique use as input the captured scene by a RGB or RGBD camera, for 2D or 3D location estimation, respectively.

Even though local evidence from image patches is important for identifying body parts, a coherent final pose estimate requires kinematic knowledge, due to the degrees of freedom on human articulations and self or external occlusion. HPE heavily relies not only on input data, but also on contextual information. The need for context led to a shift from local body part detectors, followed by spatial reasoning, to strong context-aware detectors, such as Convolutional Neural Networks (CNN) (NEWELL; YANG; DENG, 2016). These networks are the main drivers behind latest advancements, not only on HPE, but in the computer vision field as a whole.

CNNs, also called ConvNets (SIMONYAN; ZISSERMAN, 2014), are collections of layers that extract an hierarchy of features from images, through a set of filters optimized with supervised learning. In summary, a standard network is composed of convolutional layers that apply a convolution operation over every input image channel, followed by a non-linear transformation, usually ReLU (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). These convolutional layers may also be interleaved with pooling (downsampling) layers, in order to reduce memory consumption.

HPE techniques can be classified under two different main categories: Bottom-up or Top-down. Top-down techniques usually consist of tracking each human subject on the image, applying a pose estimation method for detecting each tracked person body parts and estimating each person pose individually. On Bottom-up strategies, on the other hand, each body part in a given image is detected. With this, a dense graph of all body parts on the image is created. Afterwards, a given graph reducing technique is conducted in order to reduce the graph into a set of subgraphs that encode correct poses. Finally, contextual knowledge is applied for reducing the complexity of reducing the dense graphs. In Figure 6, both HPE strategies are illustrated. While Top-down approaches are more common, bottom-up methods so far generate faster results (CAO et al., 2016). Despite the differences on both approaches, the two of them use deep learning models on the body part detection step and, among both Bottom-up or Top-down strategies, CNN's became a standard (INSAFUTDINOV et al., 2016). From this point onwards, the set of body-parts that compose an individual tracked human pose, will be referred to as skeleton.



(a) Top-down strategy. First, each human is tracked in the image and a single-person pose estimation is applied for extracting each individual body pose.

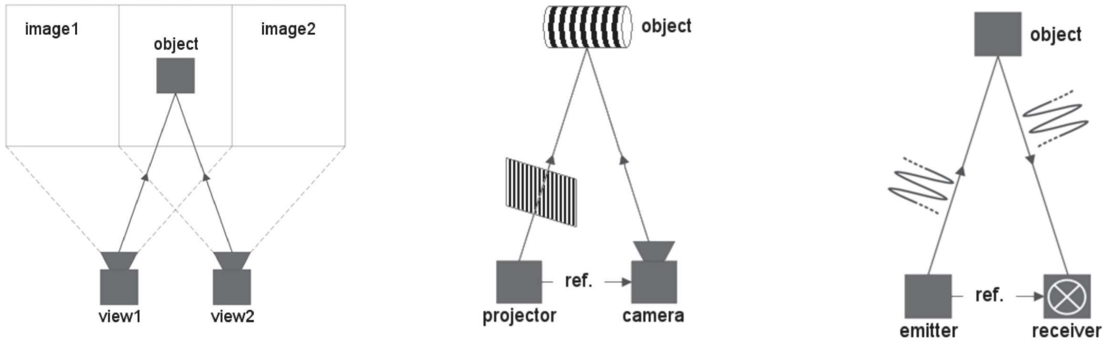
(b) Bottom-up strategy. This approach starts by detecting all body parts in the input image. Afterwards a dense graph is created and reduced afterwards. Context information is used in order to reduce the graph into a set of possible body-poses.

Figure 6: Illustrations depicting the two different generic HPE approaches.

2.2.2 RGBD Cameras

Recently, RGBD cameras became affordable and therefore had a popularity increase. RGBD cameras are devices that provide both color and depth data, enabling tridimensional readings from cameras. Such data is used as input for 3D HPE, therefore techniques as well as devices that contain such technology will be reviewed in this subsection in order to obtain better quality data as input for the HPE. Much of the RGBD devices popularity came from gaming, which is also one of the great areas of interest of HPE techniques. Microsoft Kinect's v1 was a huge success, popularizing RGBD in gaming with pose tracking. While many cameras exist, not all of them use the same principle in order to extract it's depth information. The three main principles used on RGBD devices to the current date are: Stereo Vision, Structured Light and Time of Flight. These principles will be detailed next

Stereo Vision (SV) devices work analogously to human binocular cues. Cameras that work on the top of the SV principle have two or more RGB cameras with huge overlapping FoV. After reading a frame from each RGB camera, SV algorithms are used on both read images and generate depth data. SV devices are completely passive, that is, they don't require any kind of energy emission. While this approach has several benefits, such as working on outdoor environments and not suffering from mutual interference, they are computationally expensive and unable to perform correctly on dark environments. Therefore, many times SV devices work on lower frame rate and with reduced depth size for compensating it's huge computational



(a) Stereo Vision

(b) Structured Light

(c) Time of Flight

Figure 7: Working principles of RGBD cameras. Stereo Vision (a) consists of extracting the depth by correlating data from two or more RGB lenses. (b) Structured Light projects a light pattern into the scene, and also extracts depth data by correlating the distortion from the light pattern. Time-of-Flight (c) emits modulated light pulses and estimates depth information by calculating the time the wave takes to come back. Source: Adapted from (ANTUNES et al., 2018).

needs (HUSSMANN; RINGBECK; HAGEBEUKER, 2008).

Structured Light (SL) camera devices work on a similar fashion to SV. Depth information is also extracted by correlating data on SL devices. While SV devices search for similar features on two or more images for generating depth data, SL cameras project light patterns on the camera FoV (SCHMALZ et al., 2012). With this, SL devices do not need two or more RGB cameras on the same device. After projecting the pattern onto the scene, the depth information is calculated from the deformations observed on the pattern (SCHMALZ et al., 2012; HARTMANN; SCHLAEFER, 2013). This active approach addresses several problems existing in SV. It works properly with dark scenes and depth data is computationally lightweight. However, other problems are introduced, such as: increased hardware complexity and mutual interference with other infrared-enabled depth cameras (SEEWALD et al., 2018).

Finally, Time of Flight (ToF) enabled camera devices work under the same principle as the ToA strategy for estimating location of RF devices. ToF-enabled devices are composed of an image sensor and a light emitter. The emitter emits light waves onto the scene and calculates the time it takes to come back for each pixel based on the wave speed (BAUER et al., 2013). While ToF devices also suffer from mutual interference, they are immune to lighting conditions. Further, this principle for RGBD cameras completely turns away from correspondence problems that arise on SL and SV. Lastly, ToF-enabled devices are very accurate. Suffering from variations of only 10 milimeters under certain circumstances (LACHAT et al., 2015; WASENMÜLLER; STRICKER, 2016). In Figure 7 the three described RGBD camera principles are presented.

2.3 Sensor Fusion

Sensor, multisensor or data fusion is a research area that encompasses a set of methodologies, algorithms and technologies used for combining data from different sources, heterogeneous or homogeneous, aiming at extracting a most accurate information about a specific scenario (FORTINO et al., 2019). The idea of data fusion originates in the idea that the most diverse and abundant the information sources, most robust and accurate the final information is (PAU, 1988). The application of data fusion is hardly new. With strong foundation on defense-based applications (DRAZOVICH, 1983) and with quick spread to other non-military areas such as medical diagnosis and machinery monitoring in less than a decade, data fusion started receiving significant attention in the 80s, 90s and up to today. During that period, extensive research was conducted in this area due to the emergence of new sensors and incresingly processing capabilities, which enabled real-time data fusion.

In the 2000s, the advent of Wireless Sensor Networks (WSN) brought upon a new era. Advancements on wireless communications and digital electronics enabled the development of low-cost, low-power, multifunctional small sensor nodes (AKYILDIZ et al., 2002). Such sensors could be deployed without pre-determined or engineered positions, ensuring easier and faster use. The increased accessibility to such devices further encouraged research on the area moving research in military, environmental, health and smarthome applications. This scenario stimulated the development of Body Sensor Networks (BSNs) or Body Area Networks (BANs) (CHEN et al., 2011; LATRÉ et al., 2011), used mainly for healthcare and Quality of Life improvement. Furthermore, WSNs enabled the Internet of Things (IoT) (ATZORI; IERA; MORABITO, 2010) paradigm to quickly gain notoriety and bringing forth and sustaining the envisioned Ubiquitous Computing (WEISER, 1991) idea. In the big data (MCAFEE et al., 2012; WU et al., 2013) age, together with the Internet of Things, BSNs, and the approach of large-scale use of cyber-physical systems (LEE; BAGHERI; KAO, 2015), leave us with a plethora of sensors and user generated data. According to GRAVINA et al. (2017), the combination of data fusion of multiple homogeneous or heterogeneous sensors offer several advantages, such as:

- *Improved signal to noise ratio*: fusing different sensor data reduces noisy data;
- *Reduced uncertainty*: multiple data sources reduce output uncertainty;
- *Increased confidence*: in contrast with individual sensor data, using multiple sensors increase the data reliability;
- *Enhanced robustness*: multiple sensors provide redundancy, which enhance system robustness and tolerance;
- *Improved precision*: by fusing independent measures of the same attribute, better resolution can be obtained. This particular advantage is the one to be explored in this thesis.

3 RELATED WORK

The task of tracking persons has a long history in research. And, throughout the literature, several tracking solutions are proposed with many different sensor types, with video and radiofrequency based solutions being the most explored ones. Video and radio-based solutions are pretty much complementary to each other. Video-based applications enable unobtrusive tracking of people with high frequency. Further, such solutions gained popularity in the latest years where the advancement on processing capabilities boosted CV-based tracking applications performance, and enabled a larger plethora of information to be extracted from each frame. However, CV-based applications lack the ability to consistently maintain tracked subjects information even on marker-based approaches. Radio-based applications on the other hand, are obtrusive solutions that require tracked subjects to carry a RF tag. This models consistently read the target's unique id and position if the tag is under the designed coverage area.

Considering the complementary nature of both tracking models, it is only natural to combine them. However, only a few works that fuse CV and RF-based devices exist on the literature. In the remainder of this chapter, works that conduct fusion of vision and radio sensors will be reviewed.

At a first moment, in order to guarantee revision of only relevant and related works, a set of keywords were defined. The set of keywords reflect the main technologies (*radiofrequency, vision*) the model is envisioned to work on the top of, techniques (*sensor, fusion*) and operation (*tracking*) it is expected to conduct. With this in mind, the following set of keywords was defined in order to search for literature corpus:

$$sensor \wedge fusion \wedge radio \wedge vision \wedge tracking$$

While short, this set of keywords provides a broad scope of what the present work is about. Next, this keyword set was used on Google Scholar and the most relevant results were obtained.

3.1 State-of-the-Art Review

Automatically locating people through camera's feeds is a recurring theme on the CV field. Indoor tracking with RF sensors was, for years, an open research challenge. Current solutions reached a mature state state (DARDARI; CLOSAS; DJURIC, 2015), with several industry solutions and applications currently available. With the advancements on Body Sensor Networks (BSN), the fusion of sensor-generated data became essential. BSNs can be implemented in different fashions, using homogeneous or heterogeneous data-sources. Furthermore, few studies explore the fusion of RF and CV sensors data, even if these types of devices are widely used in different applications (GRAVINA et al., 2017).

In MANDELJC et al. (2012) work, the authors track humans positions in cameras' feeds and also use a UWB RTLS solution for tracking the persons' location and identification. As a result, they fuse both cameras and RTLS data. However, the tracked humans are represented in a bounding box fashion, differently to the current proposed model, which obtains the pose for each tracked human. Another work by the same authors (PERŠ et al., 2011) proposes an alteration to the POM algorithm. The modification consists of fusing radiofrequency data into the POM algorithm to enhance its localization results.

PAPAIOANNOU et al. (2014) propose the RAVEL (Radio And Vision Enhanced Localization) positioning system. Their system also fuses RTLS and video data. The developed system is evaluated in a museum to test it against real-world environments, with constant occlusions and dim lighting. More recently, the same authors conduct experiments of tracking people on construction sites on a similar fashion (PAPAIOANNOU; MARKHAM; TRIGONI, 2017). However, in this last work, the authors add inertial measurements to their tracking parameters, resulting in increased accuracy. Finally, the authors compare their solution performance against the already proposed RAVEL model. However, once again, the obtained people visual data is a bounding box, aggregating small value to the final result. In turn, the Baptizo model enables the employment of activity recognition techniques on generated data.

TENG et al. (2014) propose an alternative approach to fuse radiofrequency and visual signals. On this work, the authors integrate electronic radiofrequency and visual signals for accurately locating and tracking nodes. The authors named their solution EV-Loc. And on the work of LI et al. (2013), the authors propose an extension of the EV-Loc model called EV-Human. The authors detect interference of human bodies on RF signals and try to correct the interference using cameras, aiming at increasing localization accuracy.

Focusing in accurately tracking consumer activities in retail environments, STURARI et al. (2016) developed a position estimation framework based on RGBD and beacons data fusion. This work focuses on reducing the error on estimated positions through data fusion. It achieves this goal with a Shopper Analytics (LICIOTTI et al., 2014) tracking technique, which uses the RGBD cameras' feed and returns a single coordinate point representing the subject's position. Experiments in a retail environment report sensible position accuracy enhancements.

Table 2: Related works key aspects.

	MANDELJC et al. (2012)	**PAPAIOANNOU et al. (2014); PAPAIOANNOU; MARKHAM; TRIGONI (2017)**	**TENG et al. (2014)**	**STURARI et al. (2016)**
RF Technology	UWB	WiFi	WiFi	BLE
CV Tracking Technique	POM (FLEURET et al., 2008)	MOG-based Background Subtraction (STAUFFER; GRIMSON, 1999)	HOG Pedestrian Detector (DALAL; TRIGGS, 2005)	Shopper Analytics (LICIOTTI et al., 2014)
Number of Markers	One (RTLS Tag)	Two (RTLS Tag and Colored Hats)	One (Smartphone)	One (Smartphone)
Output Data	Identified Bounding Box	Identified Bounding Box	Object's Location	Customer's Position

Table 2 summarizes the presented related works. It categorizes the studies according to key aspects such as the employed RF technology, CV techniques, number of markers, and output data. As shown in the table, current works propose the fusion of RF and CV data, but none of them explore the fusion of pose and RF data while also reducing RF tag position estimations, which is where the present work is focused.

4 BAPTIZO MODEL

The proposed model has two main objectives, improve the RTLS precision by fusing it's data with HPE generated skeletons and provide reliable identification data to generated skeletons. The improvement on the RTLS precision is expected to be achieved by conducting a second layer of filtering on the tags estimated location data together with the HPE data. This second filter is referenced to as Ghosting from this point on. The Ghosting heuristic is expected to further correct the RTLS position, drawing the tag even closer to it's respective skeleton, providing reliable identification data. The name of the technique originates from the idea of using a *ghost* from a pose from past frames in order to enhance the tag position estimation. This chapter is structured as follows: in Section 4.1 the design of the model is explained, detailing techniques and technologies that the model is envisioned to work on the top of; and in Chapter 4.2 the Baptizo architecture will be thoroughly explained, together with it's modules, algorithms and data structures.

4.1 Design Decision

This subsection defines the design decisions of the envisioned model. As already pointed out, the here presented model combines the data from markerless HPE generated human poses and RF devices. The proposed strategy is modelled for working on constrained scenarios, requiring that tracked subjects carry only a RF tag on a position that will be further described. In Figure 8, a generic scene with the used devices and expected deployment of the proposed model is depicted. In the scene a person to be tracked is carrying a RF tag which constantly emits it's signal which is collected by the anchor's attached on the room's ceiling, a set of RGBD cameras can be deployed throughout the coverage of the RTLS solution for tracking people's poses. The model is envisioned in a manner that the users to be tracked must only know the specific body part where to attach the RTLS tag.

Regarding the RGBD cameras. As found throughout literature and detailed in Chapter 2. While some RF technologies enable near point-level accuracy tracking, they are not as accurate as some RGBD cameras. Therefore the envisioned model is built on top of the idea that some RGBD devices yield better precision than RTLS devices. The RTLS works as a supportive technology for identifying tracked people. Any amount of RGBD cameras can be deployed with the solution, and must be placed accordingly to the needs of the HPE tool that will be used. Since the HPE tool is most likely a CNN, this model do not comprehend the network creation or training, it assumes it to be pre-trained. Therefore the cameras must be placed accordingly for returning RGBD images that match inputs that the CNN was trained on, be it placed on the ground or in the roof. Any amount of the same kind of cameras can be used.

It is important that the used RF sensors are proper for point level accuracy tracking. The RF technology will most likely consist of a set of tags and anchors, and there must be one tag for

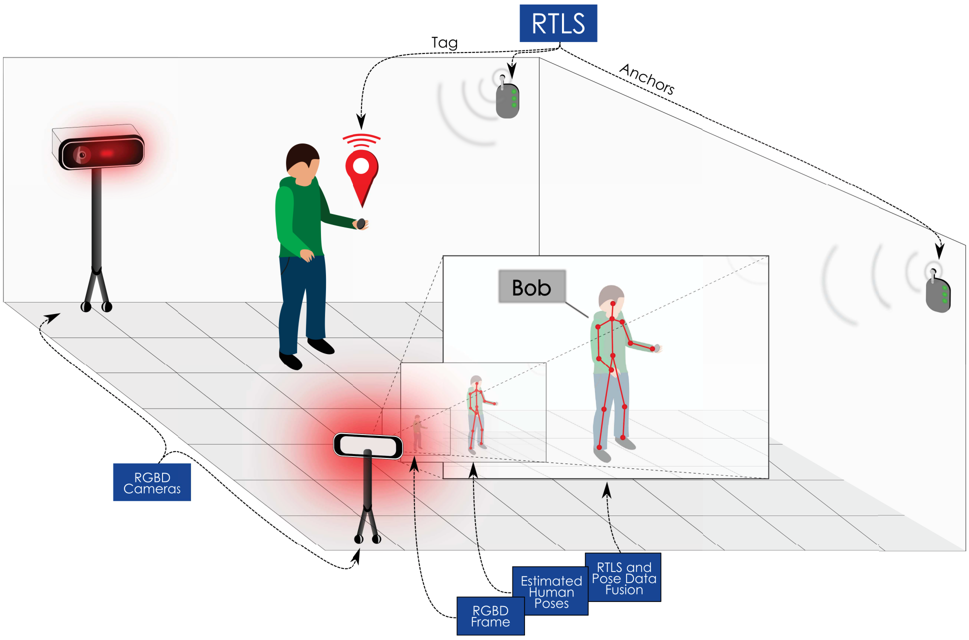


Figure 8: Scene representing the expected deployment of the model. An area is covered by RTLS anchors and RGBD cameras. People carrying tags are tracked by the RTLS and their poses are estimated using the RGBD data. Finally, both individual RTLS and Pose data are fused generating a single pose data enhanced with the tag identification.

each tracked person. Further, the tag must be attached to a predefined body part of the person. That body part must be one of the joints recognized by the employed HPE technique. This is of vital importance for this model considering that parts of the sensor fusion heuristics take into account the distance between the estimated tag location and the joint that it is attached to. From this point on, the hpe skeleton joint analogous to the body part which the RF tag is attached to will be referred to as *key joint*. Lastly, the RTLS coordinate system is the be the world coordinate system in this proposed model. Therefore all RGBD data must be transformed to the RTLS coordinate system.

The presented architecture is part of a larger project, on which other members also contributed to the work. In order to better contextualize and explain the scope of the work, the whole projects architecture will be detailed. However, the work performed on this particular thesis is restricted to a smaller subset of the model's architecure, that is the Real-Time Location System and Sensor Fusion Modules, which will be detailed on the remainder of this section.

4.2 Architecture

The proposed Baptizo architecture is composed of two different groups: the *Sensors Group* and the *Data Processing Group*. The *Sensors Group* contains all sensors used for producing required data, that is, RTLS tags and RGBD cameras. Therefore, the *Sensors Group*, in relation to this given model, is composed of two different modules: *RTLS Sensors* and *RGDB Devices*.

The *Data Processing Group* is responsible for collecting, processing, filtering, and fusing

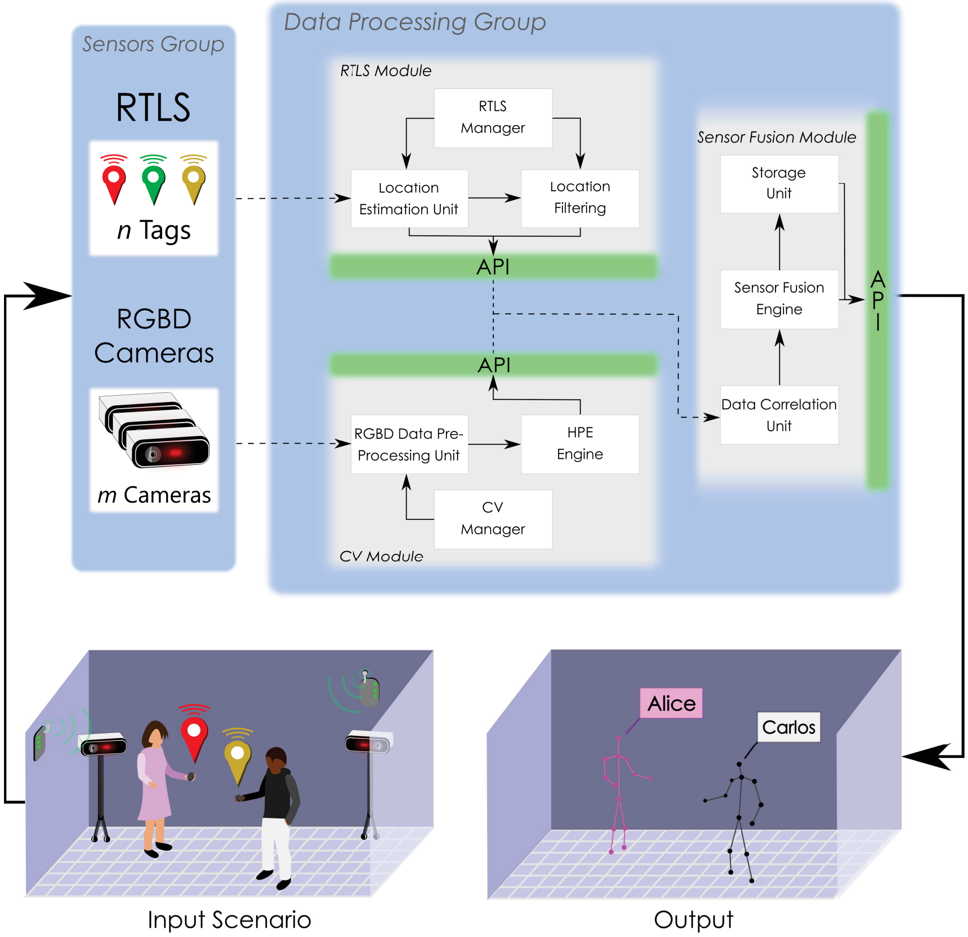


Figure 9: Baptizo architecture depicting input scenario, sensors, modules, data flow and output.

all the data produced by the sensors. This group is composed of three different modules: *RTLS Module*, *CV Module* and the *Sensor Fusion Module*. Each module of the *Data Processing Group* is self-titled according to the data or procedure it is expected to compute. Both *RTLS Module* and *CV Module* have a management component that contains a set of parameters and tools that are meant to process their data. These parameters will be further detailed on the remaining of this chapter. The *RTLS Module* consumes the RF tags signals data collected by the anchors, estimates, and filters each tag position, providing position data through an Application Programming Interface (API). Working in a similar fashion, the *CV Module* uses the RGBD data produced by the cameras and estimates the poses of people on the Field-of-View (FoV) of the camera, also providing generated data via API. Finally, the *Sensor Fusion Module* consumes the computed tag locations and poses by the beforementioned modules and fuses their data, enhancing the tag locations precision and giving an identity to data from the HPE module. The APIs enable running each module in different computing nodes, depending on infrastructural needs. These APIs are developed in a publish-subscribe fashion, and are projected only to be used for transferring data between the architecture modules.

Figure 9 illustrates the Baptizo model architecture. The arrows show the data flow direction, dashed lines represent intermodular data flow, and solid lines represent intramodular data flow.

Gray boxes inside the *Data Processing Group* represent its internal modules, and the white boxes inside these modules represent its components.

An overview of the dataflow and generated data of each module of the *Data Processing Group* is presented in Figure 10. The CV and RTLS modules consume pre-processed data, namely RGBD frames and unfiltered estimated tag positions. The CV Module applies a 3D HPE technique on the RGBD frames generating *n* 3D skeletons, and the RTLS Module applies *f* filters on the *t* tags data and generates *f* filtered tag positions for each *t* tag. Lastly, the Sensor Fusion receives both data and generates *t* enhanced location data, composed of with *n* identified 3D skeletons.

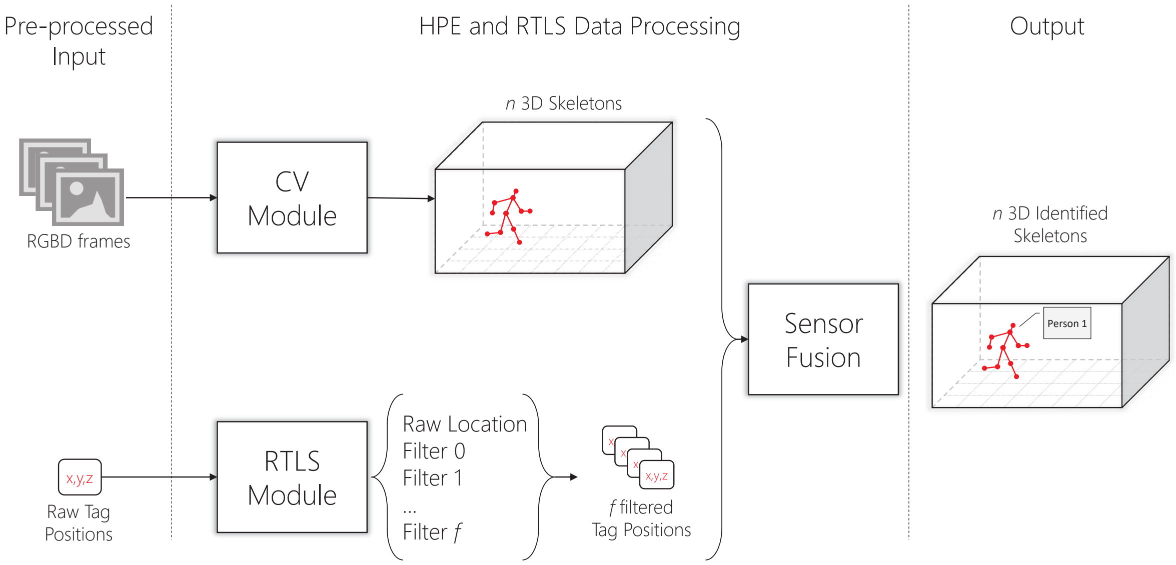


Figure 10: HPE and RTLS Dataflow.

4.2.1 RTLS and CV Modules

The *RTLS Module* is responsible for estimating the RF tags positions, applying filters and returning a set of *x, y, z* coordinates for each tag reading. The module has four components: the RTLS Manager; the API; the Location Estimation Unit; and Location Filtering Unit. The Manager sets the module parameters of the reading. These parameters consist of a set the techniques or filters specific for each reading, as well as a *FrameCount* and *Timestamp* for synchronization with the CV counterpart module. For each tag reading, the Location Estimation and Filtering Units conduct their operations accordingly to the parameters. This way, different amounts of information may be obtained by applying different localization techniques and filtering on the readings.

The CV module consumes the RGBD cameras data as input and provides a set of poses as output. Estimated poses consist of a set of coordinates for each body part recognized by

Table 3: RTLS Manager Parameters.

Parameters	Description
`Type`	Data type for the Sensor Fusion module interpret and handle data accordingly.
`Localization Techniques`	Localization technique to be used on estimating the sensors location.
`Filters`	Set of f filters to be used.
`FrameCount`	Sequential number of readings.
`Timestamp`	Time in milliseconds of the data collection.

the HPE. This module also has four components: the CV Manager; the API; the Data Pre-Processing Unit; and the HPE Engine. Analogously to the RTLS counterpart, the CV Manager also has a set of parameters that are applied and used on each of the module's operations, such as *FrameCount*, *Timestamp* and *Transformation Matrices* used for transforming the cameras' coordinate spaces into the world's. The Pre-Processing Unit is responsible for transforming all cameras' coordinate spaces into a singular one and pre-process the data for inputting into the HPE Engine. The HPE Engine applies any HPE technique that consumes the pre-processed RGBD data and returns sets of body parts composing poses. In Table 4, all parameters are presented.

Table 4: CV Manager Parameters.

Parameters	Description
`Type`	Data type for the Sensor Fusion module interpret and handle data accordingly.
`Transformation Matrices`	Set of transformation matrices for each RGBD device.
`FrameCount`	Sequential number of readings.
`Timestamp`	Time in milliseconds of the data collection.

Figure 10 illustrates the operation of both modules. At the start of each reading of the CV Module, it collects all available RGBD frames from the devices and sets the *Timestamp* and *FrameCount* of the retrieved data. With all data pre-processed, the module advances to the HPE Engine module. The HPE uses the pre-processed RGBD images as input and returns a set of skeletons, which are groups of coordinate points representing the joints of each detected body part. Finally, in case that there is more than one camera, the best-estimated poses must be matched. With this, the module provides a group of poses with the body parts estimated with higher confidence. The RTLS Module, on the other hand, collects the tag signals from the anchors and, using a given localization technique, estimates the tag position. Next, a filtering technique further processes the resulting positions. This filter is a derivative of the Kalman method (KAUTZ; GROH; ESKOFIER, 2016). Finally, the module outputs a set of coordinate points for each tag reading.

4.2.2 Sensor Fusion Module

The sensor fusion module is the core of the proposed model. In this module, the unidentified skeletons generated by the HPE are fused with the RTLS tags unique ids and positions.

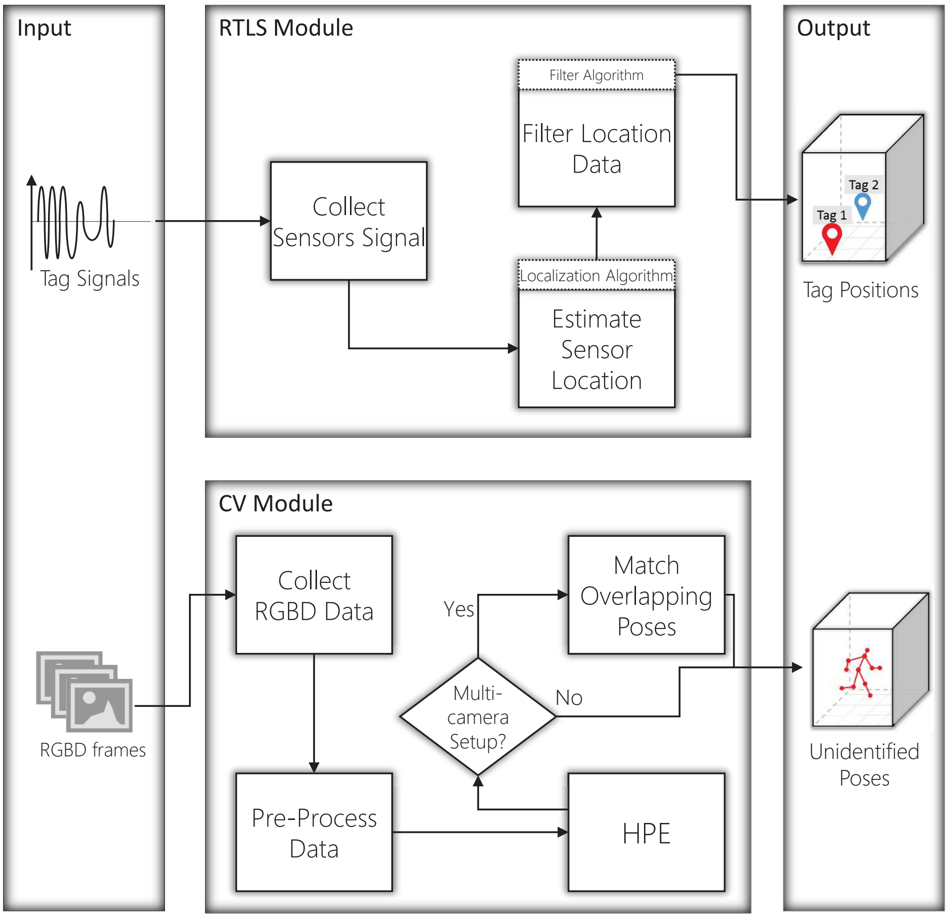


Figure 11: RTLS and CV modules flowchart with inputs and outputs. The RTLS Module consumes the tag signals, estimates and filters its position, and returns it as coordinate points. The CV Module receives all RGBD data from the cameras, and transforms all coordinates into the world's coordinate space before estimating the human poses and outputting the estimated human poses.

By consequence of the proposed model, the module outputs a generated set of identified human poses. The module has four components: the Data Correlation Unit, which is responsible for synchronizing consumed RTLS and CV data; the Sensor Fusion Engine, responsible for applying heuristics, fusing the data; a Storage Unit for storing output data; finally, an API provides stored data and enables other applications to subscribe to the module's output data feed. At first, the component retrieves the data of both RTLS and CV modules and their data correlated according to the respective timestamp. The procedure requires all nodes running the RTLS and CV modules to be synchronized. The Sensor Fusion Engine receives packages of tag positions and poses, and proceeds to fuse the data and apply heuristics for improving the fusion quality. The Engine consists of three main steps: (i) the fusion process itself, which returns a `Map` of tags and skeletons; (ii) the initiation process, which is repeated until a stable scenario is met; (iii) the main sensor fusion procedure with heuristics, step focused on applying the Ghosting technique. Therefore, there are two categories of data fusion results generated by the proposed model. The first fuses pose data with RTLS tag identifications. In turn, the second aims at reducing the tags position estimation error by filtering that data with most-accurate depth data from the RGBD cameras. The used strategies and algorithms are the subject of the following discussions.

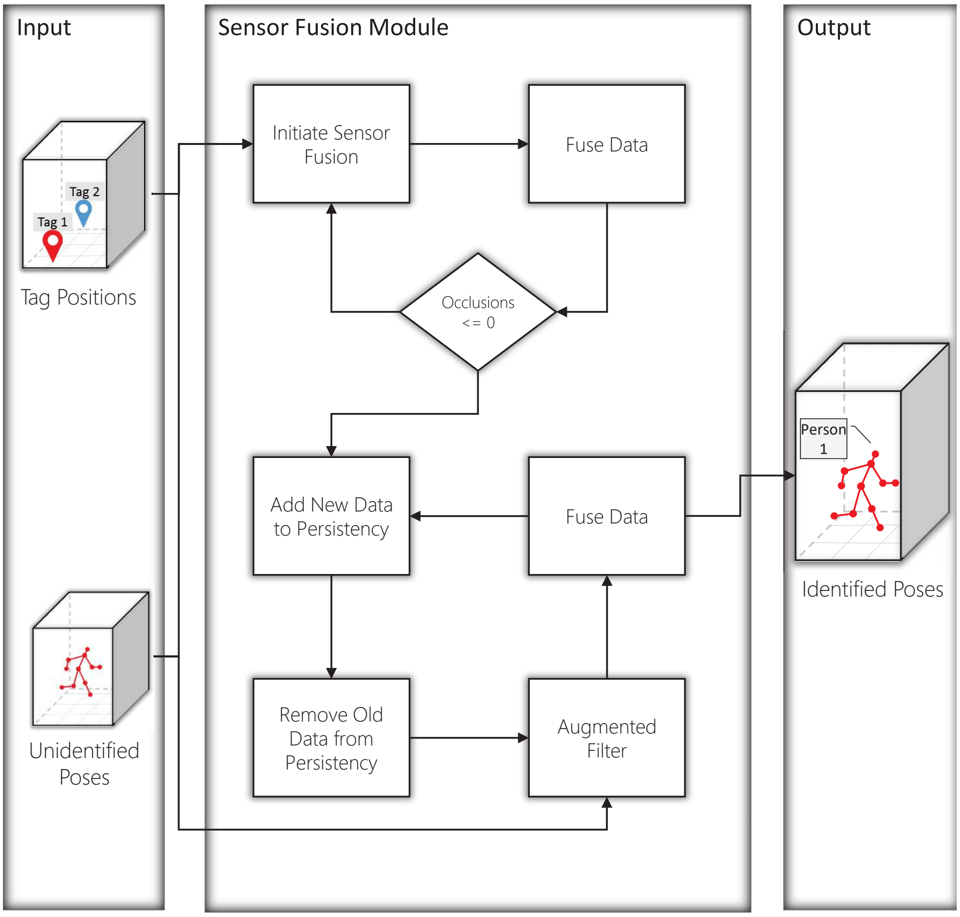


Figure 12: Sensor Fusion flowchart. The Sensor Fusion Module consumes both CV and RTLS modules's data. In a first moment, the module loops through series of fusions and conditions. This process consists of a sequence of data fusions based on tags and poses euclidean distances. Once no occlusions are detected, that is, there is at least one estimated pose per tag, the module proceeds to its main heuristics, managing a data persistency and using a filter that uses fused data as *a priori* states in order to reduce RTLS measurements error.

A top-view flowchart of the Sensor Fusion Module is presented in Figure 12. First, the Sensor Fusion Module consumes the data obtained from the CV and RTLS components. Then, it goes through a loop of fusions and conditions trying to reach a minimal stable state. This minimal stable state is the one in which at least one pose exists for each tag, that is, the number of occlusions is equal or less than zero. Once reaching the stable state, the algorithm considers that there are no occlusions and proceeds to the main cycle where the persistence of fused data is managed. The goal of persistence is to reuse previous iterations of successful fusions of tags and poses. It employs the key joints of persisted poses as an *a priori* state for a given filter to further correct the tag position estimation. This methodology enables the correction of tag position using accurate joint data, further approximating the result to the ground-truth position and increasing the chances of fusing it with the correct pose in the next frames. The strategy for reducing the RTLS estimation error with past fused frames skeletons is baptized as Ghosting, or Augmented Filter. All the data that goes through the augmented filter is fused, returned, updated on the persistence and passed on to the next sensor fusion iterations.

This flowchart can be translated into three different algorithms, each one responsible for different parts of the proposed fusion strategy. First and foremost, a generic fusion strategy referred to as `FUSE` receives the current tag and skeleton readings and iterates over each component calculating the distance from each tag to each skeleton, then fusing the tags which are closer to each skeleton. This function is called by the other two algorithms for fusing the read RTLS and HPE data. The `INIT` algorithm is the first loop discussed on the previous flowchart. This algorithm can be explained as a sequence of readings and calls to the `FUSE` algorithm trying to reach the previously discussed stable state. Once the stable state is reached, the algorithm proceeds to the main `SENSOR_FUSION` procedure. This main procedure is responsible for managing the discussed data persistance and updating the current RTLS readings with the proposed filter using the HPE data as *a priori* data, and then passing the updated tag data package to the fusion algorithm. Next, these three algorithms will be thoroughly discussed and detailed.

The fusion procedure is presented in Algorithm 1. This procedure is used during the whole fusion process. It receives the synchronized tag and skeleton packages as parameters and iterates over the skeleton package. For each skeleton `S` in the package, the algorithm searches for the closest tag `closestTag` in the tagPackage to the skeleton `S` key joint and inserts it into a fused data map `dataMap` with the closest tag as the *key* and the skeleton `S` as *value*. The `closestTag` is removed from the tagPackage. Next, for each remaining tag `T` in the tagPackage, `T` is inserted with no value on `dataMap`. Finally, `dataMap` is returned.

Algorithm 1 Fuse Pseudocode

```
procedure FUSE(tagPackage, skeletonPackage)
    $dataMap\langle\rangle \leftarrow new\ Map\langle Tag, Skeleton\rangle$
    for each skeleton $S$ in $skeletonPackage$ do
        $closestTag \leftarrow$ FIND_CLOSEST_TAG($tagPackage, S.keyJoint$)
        $closestTag.position \leftarrow S.keyJoint.position$
        $dataMap.put(closestTag, s)$
        $tagPackage.remove(closestTag)$
    end for
    for each remainingTag $T$ in $tagPackage$ do
        $dataMap.put(T, \varnothing)$
    end for
    return $dataMap\langle\rangle$
end procedure
```

The Init is a procedure that repeats itself until a minimally stable state is reached. A state is considered stable when at least one skeleton exists for each tag, that is, the number of occlusions is equal or less than zero. First, the procedure collects the data packages and fuse them, passing the collected data to the `fuse`. Next, the number of occluded people is calculated by subtracting the amount of tags in `tagPackage` by the number of skeletons in `skeletonPackage`, which is the output of the HPE conducted in the CV Module. According to one of the design decisions of the model, there shouldn't be more skeletons than the number of tags, therefore if the number of occluded is less than zero, `init` is returned. Otherwise, the fused data `dataMap` is stored and if the number of occluded people is zero, the Sensor Fusion proceeds to it's main procedure `sensor_fusion` passing `dataMap` as a parameter, otherwise the initiation procedure repeats itself. With this, a stable state where each tag is fused to a skeleton is guaranteed before starting the main procedure. The Init pseudocode is presented in Algorithm 2.

Algorithm 2 Init Pseudocode

```
procedure INIT
    $tagPackage, skeletonPackage \leftarrow readData()$
    $dataMap\,\langle\rangle \leftarrow$ FUSE($tagPackage, skeletonPackage$)
    $occluded \leftarrow tagPackage.size - skeletonPackage.size$
    if $occluded < 0$ then
        $warning$
        return INIT
    end if
    STORE($dataMap\,\langle\rangle$)
    if $occluded == 0$ then
        return SENSOR_FUSION($dataMap\,\langle\rangle$)
    else
        return INIT
    end if
end procedure
```

The Sensor Fusion algorithm is presented in Algorithm 3. This algorithm refers to the main phase after reaching the stable state of the sensor fusion, that is, the loops after no occlusion is detected. On this algorithm, skeletons are analogous to the so-far called poses, and the persistence mentioned above, managed throughout the entire sensor fusion process for the Ghosting heuristic, is named `dataPersistance`. The algorithm has two inputs: the first stable fusion data set θ, and the `dataPersistance`. The data persistance is a *Map* that has a *Pair* of Tag and Skeletons as *key* value and an *int* as value. The *key* value is the fused data while the *int* value is the age of the data, counted as the number of frames that that specific data is on the persistence. Every time a new occurrence of that fused data appears, it overwrites previous ones of the same tag identification. Or, if the age of the fused data on the `dataPersistance` is equal than a given threshold τ, the algorithm removes the data entry. The threshold guarantees that the current tag readings are not being drawn to a pose that is suffering from long-term occlusion, hence possibly bringing it to a wrong position. This threshold must be defined accordingly to how many FPS are being processed and according to the specific behavior of the monitored site.

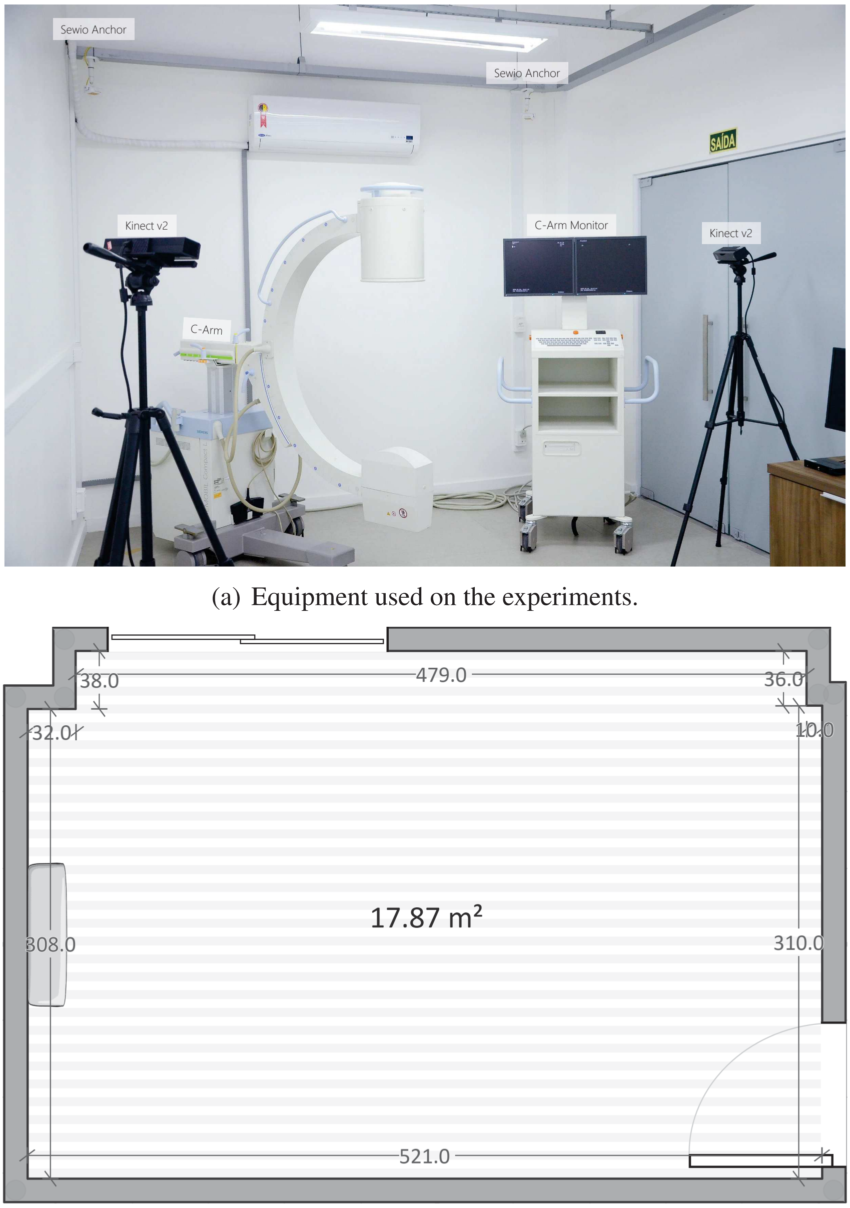
Algorithm 3 Sensor Fusion Algorithm

```
Input:
The first fused data $\theta$
A map $dataPersistance\ \langle Pair\ \langle Tag, Skeleton\rangle\ , int\rangle$
procedure SENSOR_FUSION($\theta$, $dataPersistance$)
    for each Pair $\langle Tag, Skeleton\rangle$ $data$ in $\theta$ do
        $dataPersistance.put(data, 0)$
    end for
    for each Pair $\langle Tag, \varnothing\rangle$ $data$ in $\theta$ do
        if $dataPersistance.get(data) == \tau$ then
            $dataPersistance.remove(data)$
            continue
        end if
        $dataPersistance.at(data) += 1$
    end for
    $tags$, $skeletons \leftarrow readData()$
    AUGMENTED_FILTER($tags$, $dataPersistance$)
    $\theta \leftarrow$ FUSE($tags$, $skeletons$)
    STORE($\theta$)
    return SENSOR_FUSION($\theta$, $dataPersistance$)
end procedure
Output:
Frame fused data $\theta$
```

As an output, the algorithm saves the frame fused data.

5 METHODOLOGY

In this chapter, the methodology for deploying, developing and testing the proof-of-concept of the proposed model will be thoroughly detailed. Figure 13 presents the used equipment, its deployment, and the experiment room measurements. To create a deployment scenario based on real-world use cases, e.g., an Operating Room (OR), a Siemens C-Arm and monitor are used.



(a) Equipment used on the experiments.

(b) Room measurements in centimeters.

Figure 13: Photos of equipment and measures of the experiment setup.

Regarding sensor choices, the Sewio UWB RTLS-TDoA Kit [1] was the RTLS of choice. This RTLS is UWB-based and according to the manufacturers specifications, it has sub-room accuracy levels down to 25cm of error. This solution consists of a set of five anchors which are

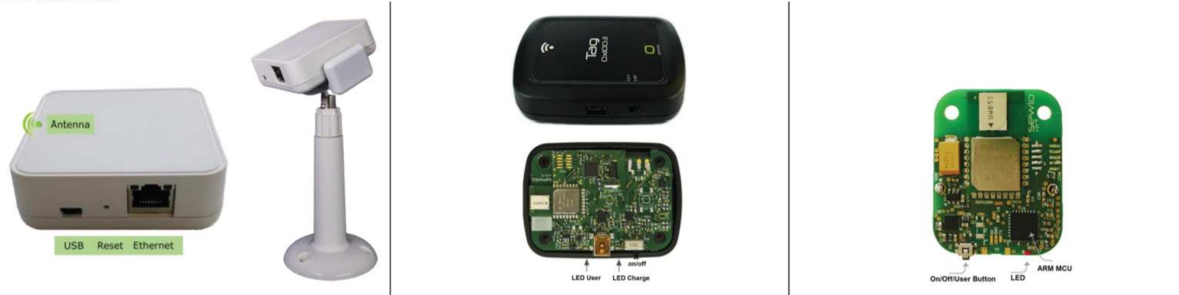
[1] https://www.sewio.net/

to be fixed in the room ceiling and UWB tags carried by the monitored subjects. The Sewio RTLS estimates tag positions using the already discussed TDoA localization technique. Each component of the Sewio Kit is depicted and described in Figure 14. It is important to note that this particular RTLS is two dimensional.

Anchors
Compliant with UWB PHY IEEE 802.15.4a
Decawave UWB Radio, 6 channels 3-7GHz
Dimensions: 70x74x25 mm
Driven by MCU ARM Cortex M4
Configurable via web RTLS Manager
Anchor's wireless sync via UWB
Ethernet used as a backhaul
Firmware upgrade via Ethernet
Native web interface
For Indoor Use

Li-ion Tags
Compliant with UWB PHY IEEE 802.15.4a
Decawave UWB Radio, 6 channels, 3-7GHz
Dimensions: 70x50x21 mm
Driven by Ultra Low Power ARM EFM32G M3
Battery included, Li-ion 600mA
Configuration via RTLS Manager
Firmware upgrade and Charging via USB
User LED and Charging LED indication
Unique 6 bytes ID

Piccolino Tag
Compliant with UWB PHY IEEE 802.15.4a
DecaWave UWB Radio, 6 channels, 3-7GHz
Dimensions: 29x37x11 mm
Driven by Ultra Low Power ARM EFM32G M3
Coin Battery CR 2450 600mA
Configuration Wirelessly via RTLS Manager
User LED indication
Unique 6 bytes ID



(a) Anchor (b) Li-ion Tag (c) Piccolino Tag

Figure 14: Sewio RTLS-TDoA (Development) Kit technical details.

Some preliminary tests evaluating Sewio's UWB RTLS solutions accuracy were conducted in order to evaluate the system's performance. Such experiments are relevant to the scope of this work, since the more accurate the RTLS is, the closer to the respective estimated skeletons the tags are expected to be. These experiments were conducted in the OR where the final prototype of the proposed model will be deployed. The methodology applied in this experiments consisted of placing a number of RTLS tags throughout the OR on fixed positions. Afterwards, the physical location of the tag, i.e. the ground-truth, was measured accordingly to the RTLS virtual coordinate system position. All UWB tags refresh rates were set to 100 ms, that is, 10 readings per second. With this, the RTLS position estimation errors and the average readings per second were evaluated in three different scenarios. It is important to note the data evaluated here is unfiltered and the observed errors can be reduced with the use of appropriate techniques, such as a Kalman Filter. The error of the readings was measured using the Euclidean distance in the plane (Equation 5.1).

$$distance = \sqrt{(x_{estimated} - x_{ground})^2 + (y_{estimated} - y_{ground})^2} \tag{5.1}$$

Regarding cameras, the RGBD sensor of choice was the Microsoft Kinect v2[2]. The Kinect v2 works accordingly to the Time-of-Flight (ToF) principle, which consists in estimating the distance for each pixel by calculating the time taken for emitted signals to reach their targets and return to the device while Sewio's UWB works using the previously discussed TDoA technique.

[2] https://developer.microsoft.com/en-us/windows/kinect

The experimental setup employs three Kinects positioned as close as possible to the room's wall, with their FoV focused at the room's center, using a predefined setup in order to optimize FoV overlap (SEEWALD et al., 2018).

For each sensor, an individual node consumes and preprocesses the sensor data. Three nodes extract Kinect's RGBD data and conduct spatial transformations, while one node executes the RTLS solution to estimate the positions of tags. A fifth node, namely the HPE node, consumes the data from all other nodes and outputs the final data. The node is composed of a Core i7-7700HQ Processor, 16 GB of RAM and a NVIDIA GeForce GTX 1060 Graphics Processing Unit (GPU). The HPE node contains the HPE and Sensor Fusion Modules and processes data at a varying three to four FPS. Given these constraints, the Ghosting heuristic uses two frames as processing threshold for experiments, according to detailed in Section 4.2.2.

The HPE node estimates people's poses using a CNN built using the Tensorflow framework (ABADI et al., 2016) and based on the MobileNetV1 architecture (HOWARD et al., 2017). Tensorflow is an open source framework for developing neural network models that are widely used in research and enterprise deep learning related projects. Additionally, to create the CNN model, a dataset of images with persons in various poses must be used to train the neural network to recognize correct postures in different scenarios. The Microsoft Common Objects in Context (MS-COCO) (LIN et al., 2014) is used for the proof-of-concept testing and evaluation. It is a CV dataset for image recognition widely used in work related to HPE. It is important to note that it is possible to enhance the CNN model by using footage obtained from environments where the monitoring architecture will be utilized. The prototype is written in C++ and Python and uses a set of libraries which are detailed in Table 5.

Table 5: Current libraries used on the prototype.

Library	**Version**	**Description**
CMake	3.11	CMake to build the Visual Studio project.
Boost	1.6.6	Set of libraries and utilities for C++.
OpenCV	3.4.1	Computer vision libraries.
Qt	5.9	Cross-platform framework for developing UI in C++.
Flatbuffers	1.9	Library for data serialization.
Kinect v2 SDK	2.0	Kinect v2 API.
CUDA	9.0	GPU-enhanced programming library.
cuDNN	7.0	GPU-enhanced deep neural network library.
GLM	0.9.8.5	Header only C++ library for graphics software.
Tensorflow	1.5	Library for neural network model development.

Initially, a proof-of-concept of the proposed identification-pose fusion process is conducted. Two subjects carrying RTLS tags on their right wrists were tracked. After pre-processing the data, conducting all needed spatial transformations and choosing the best poses from each camera, the distance from the RTLS tag to each pose is calculated. These distances are estimated by calculating the Euclidean Distance from each tag to all poses key joints.

Due to the existing estimation error on the RTLS, this simple approach of fusing the tag to the closest pose may lead to identity switch issues. For that reason, the proposed Ghosting heuristics for RTLS error reduction detailed in the in Section 4.2.2 is proposed. It consists of saving the latest occurrence of the joint coordinates detected by the pose estimation module where the RTLS tag is attached to for τ readings. The saved joints are used as an *a priori* state for further filtering *a posteriori* RTLS states. Every new set of joints updates the previous occurrence. As stated, this strategy is correlated to the first identity fusion quality, since the tag is filtered using the previously fused pose data.

For each new frame, the methodology uses the last saved occurrence of the joint of the pose which the tag was attached to filter the current RTLS tags positions. Figure 15 illustrates a two-frame example of the Ghosting technique. Initially, in frame 0, the Frame Skeleton and the Tag Position are separated. Their data is then fused and saved on a Persistency Map. In the next frame, the Frame Skeleton and the Tag Position are once again separated data. The current tag position is filtered using the persistence data to draw the tag closer to its respective pose and increase the chances of better fusion results, moving the tag closer to its correct position.

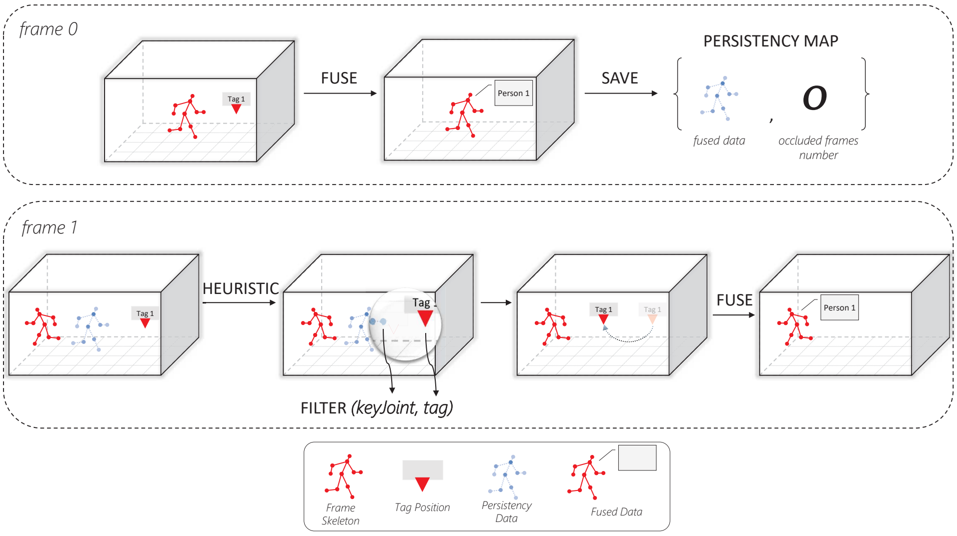


Figure 15: Ghosting heuristics. In the first frame, the respective Frame Skeleton and Tag Position are fused and added to a data persistency. On the next frame, the Frame Skeleton and the Tag Position are once again separated data. Using the Persistency Data, the Tag Position is filtered with the joint coordinates, drawing the Tag closer to its correct position, increasing the chances of a successful fuse.

Two scenarios are employed to evaluate the proposed model. The first experiment is a proof-of-concept of the proposed pose and identification fusion. It consists of two steps. In the first step, the system collects positions from RTLS tags and poses from RGBD frames, and applies calibration parameters to transform the data into a single coordinate system. Then, in the second step, the system selects the best pose occurrence for each frame and fuses the closes RTLS tag identity by verifying the closest tag to each key joint. In this experiment, two subjects were present with the RTLS tags attached to their right wrists.

The second experiment aims at analyzing the Ghosting strategy, with one and two subjects. These tests have the same approach as the pose-identification proof-of-concept, with subjects carrying the RTLS tags attached to their right wrists, which is the key joint for the filtering process. The error is calculated assuming the right wrist coordinates as estimated by the HPE body part detector as ground-truth values, considering the millimetrical accuracy of the RGBD depth data. Further, results only take in account high-confidence key joint data. These confidence values are returned by the CNN used for estimating human poses.

The proof-of-concept experiment is conducted in a experimentation lab, a controlled site where clothes, lighting and equipments are used and placed in the most favorable way aiming at obtaining nearly optimal results for proving the concept of pose-identification fusion. The Ghosting experiment is performed in two different environments, on the experimentation labs and a hybrid Operating Room (OR) located in the Instituto de Cardiologia – Fundação Universitária de Cardiologia (IC-FUC) [3]. The OR is 6,4 meters long and 5,57 meters wide and is equipped with a Siemens Axiom Artis Zee Floor. This particular room is composed of other two smaller rooms: the control room; and the equipment room where one of the servers is stored. Figure 16 presents a blueprint of the room.

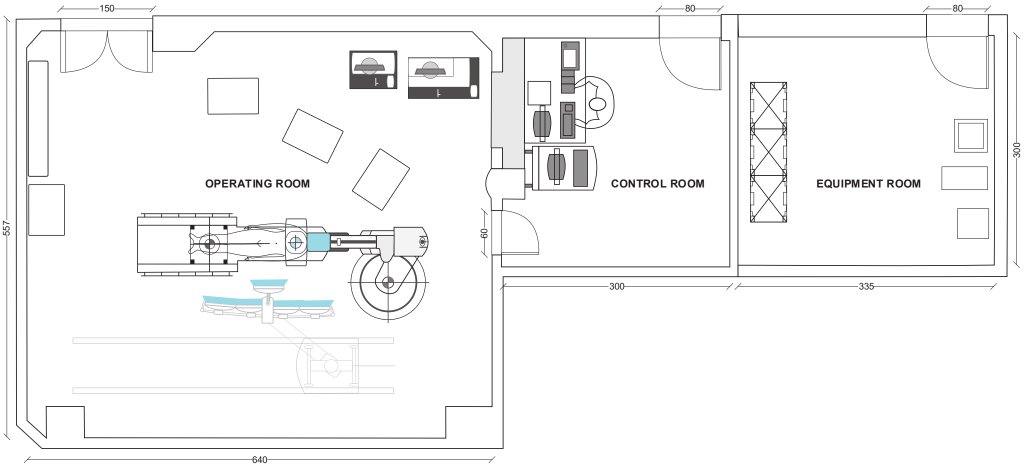


Figure 16: Operating Suite complex composed by the Operating Room on the left, a control room in the center and an equipment room on the right. Room measures in centimeters.

The list of deployed equipment on the Operating Suite (OS) is detailed in Table 6. The sensors consists of the previously detailed RGB-D and UWB sensors. A Dell OptiPlex 3050 and three Dell OptiPlex 3050 Mini are used as nodes for processing the sensors data. Lastly, network equipment was also installed on-site, consisting of a Sewio PoE Switch for powering the anchors and a Dell N1524 Switch.

The equipment installation was conducted with the supervision of the hospital's biomedical engineering team. The installation took place on a Saturday where the OR was idle and lasted about a whole day in order to ensure that all cables were invisible and all equipment was properly attached and installed. During the installation, eight UTP cables were pulled from

[3] http://www.cardiologia.org.br/

Table 6: Equipment list installed in the OR.

Equipment	Amount
Dell N1524 Switch	1
Sewio PoE Switch	1
Dell OptiPlex 3050	1
Dell OptiPlex 3050 Mini	3
Microsoft Kinect v2	3
Sewio UWB Anchor	5

each computer node and UWB anchors to the control room through its roof, and one cable pulled from the node installed on the equipment room. Both Dell and Sewio PoE switches are installed in the control room where the prototype operator will be located. The equipment distribution and network topology is presented on Figure 17.

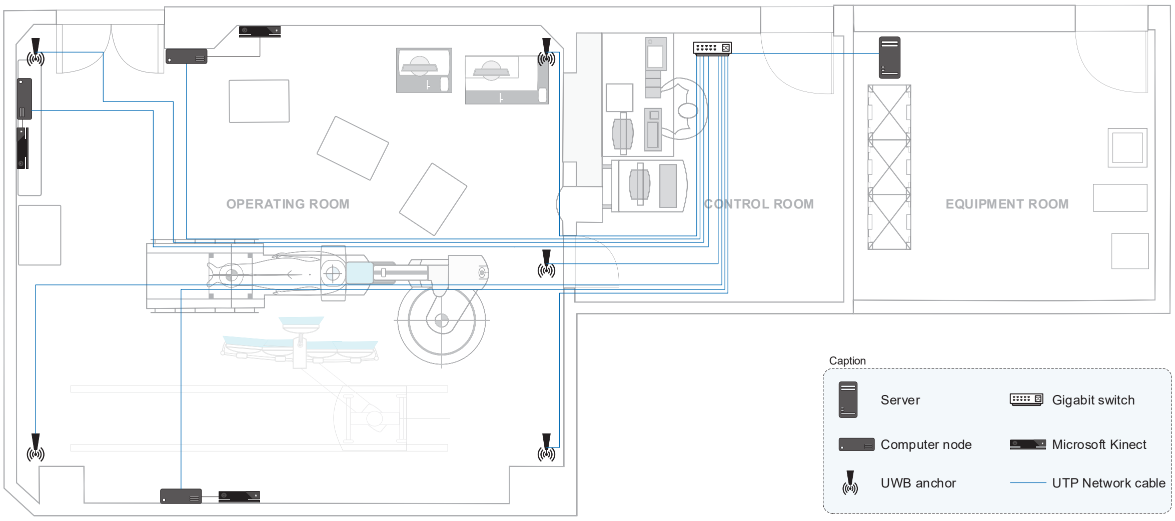
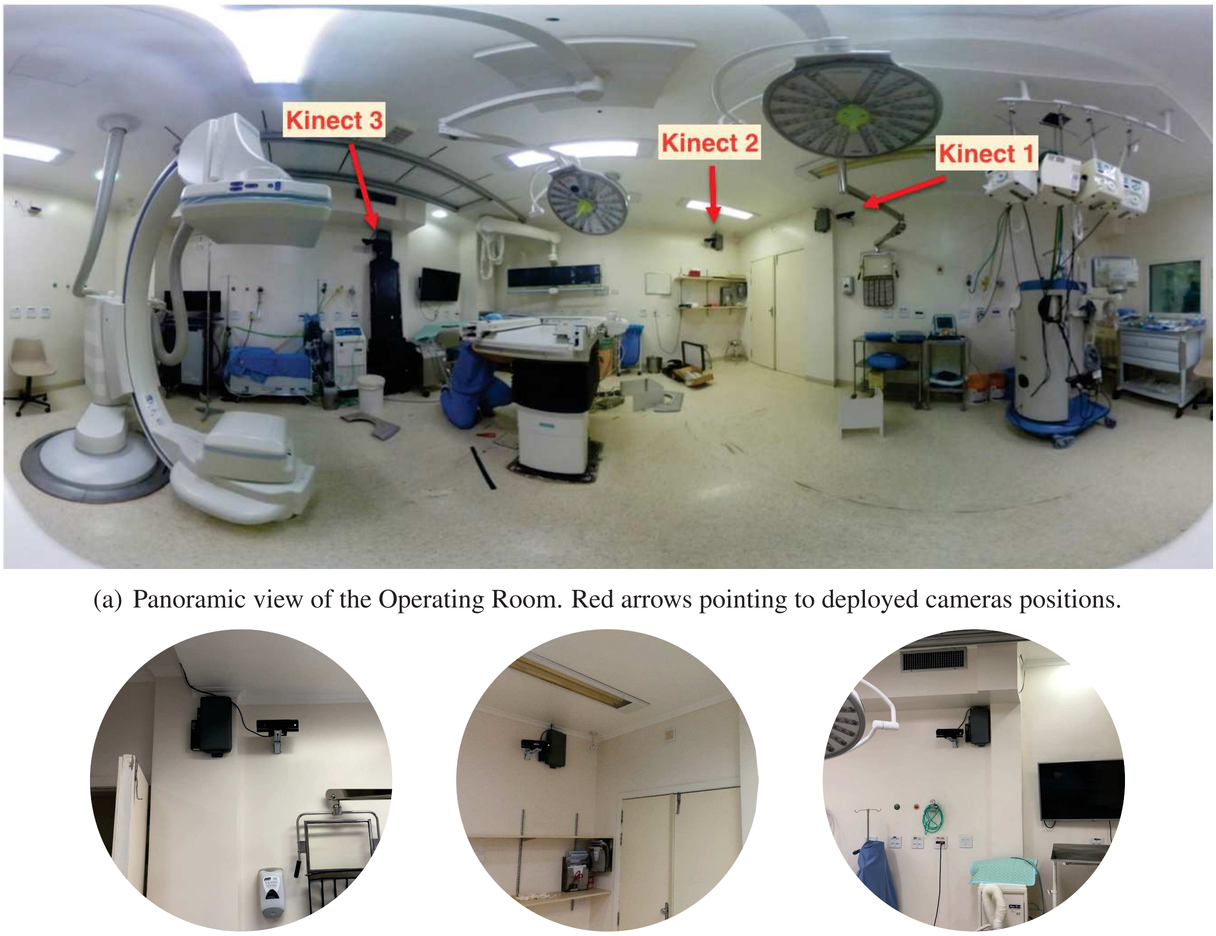


Figure 17: Equipment distribution in the Operating Suite.

At the date of this thesis, the prototype is working in an isolated private network. Contact with the Information Technology (IT) department of the hospital is underway and they are currently working in providing an external access link for us to update the prototype and retrieve data. As reported by the IT department, this process is expected to take some time due to the sensitivity of opening an external access to the hospitals infrastructure. Lastly, Figure 18 presents a panoramic view of the OR and the installed cameras.



(a) Panoramic view of the Operating Room. Red arrows pointing to deployed cameras positions.

(b) Kinect 1

(c) Kinect 2

(d) Kinect 3

Figure 18: Operating room panoramic view and deployed RGB-D cameras.

6 DISCUSSION

This section discusses the results obtained with the previously detailed methodology. Two different types of experiments are described on this section. The first focus on analyzing the deployed RTLS accuracy and observing potential interference with other medical equipment deployed on the OR. The second tests aim at analyzing the proposed model. These experiments are divided into two different categories, one for analyzing the proposed pose and RTLS data fusion. The other to analyze the error eduction when using the Ghosting heuristic, which focuses on reducing RTLS position estimation errors by using most-accurate (YANG et al., 2015; LACHAT et al., 2015) depth information of the RGB-D devices as ground-truth data for filtering of the RTLS readings.

6.1 Real-Time Location System Evaluation

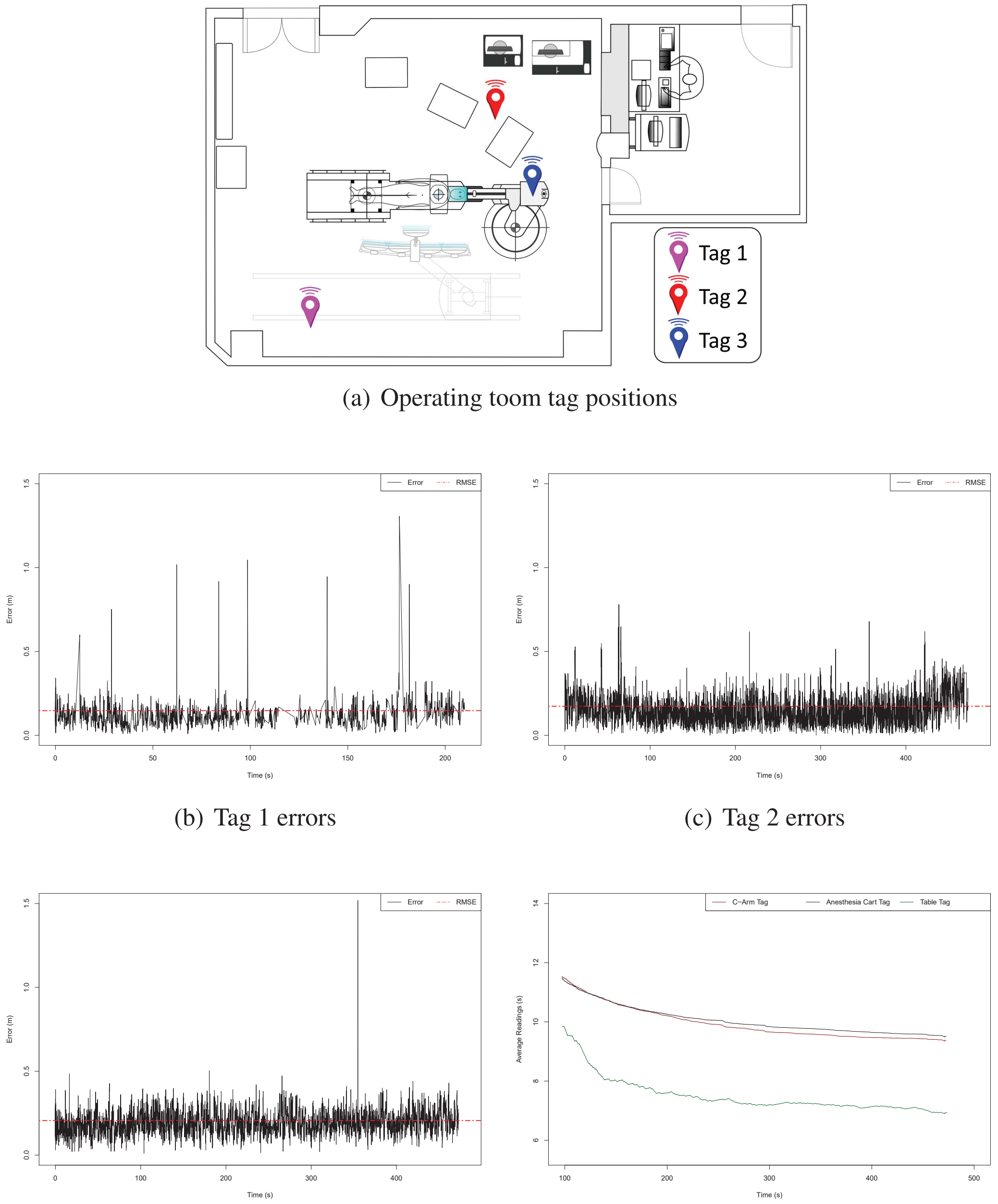
In this section, the three experiments results evaluating the UWB accuracy are detailed. Each experiment was run in a different scenario. The first experiment was conducted in an idle OR, the second during an emergency surgery and the third one during an angioplasty. The following three subsections details the results observed on each experiment.

6.1.1 Accuracy Experiments: Scenario 1 – Idle Room

The first scenario was an initial testing with the RTLS at the clinical partner where the prototype will be deployed. The aim of this experiment was to analyze if the deployed UWB anchors were working according to the manufacturer's specifications. In this experiment, three UWB tags were placed on the room and their errors estimated for approximately eight minutes.

Figure 19(a) depicts the blueprint of the room with tags' positions, the measured errors of all tags, and the average readings per second. The observed Root-Mean Square Error (RMSE), in centimeters, are of 14 for Tag 1, 17 for Tag 2 and 20 for Tag 3, as illustrated on Figures 19(b), 19(c) and 19(d). On Figure 19(b) it is possible to see that Tag 1 has larger standard deviation due have being placed close to the borders of the room.

Tag 1 has a reduced number of readings compared to the others because an accidental movement introduced unnacurate error measurements in the readings. Nevertheless, erroneous values were isolated removed from the sample. This experiment is mostly for analyzing if the RTLS precision is working accordingly to the described by the manufacturer. The observed readings, illustrated on Figure 19(e), were of 10 per second for tags 2 and 3 and of approximately 8 for tag 1, which is also a accordingly to the expected.



(a) Operating toom tag positions

(b) Tag 1 errors

(c) Tag 2 errors

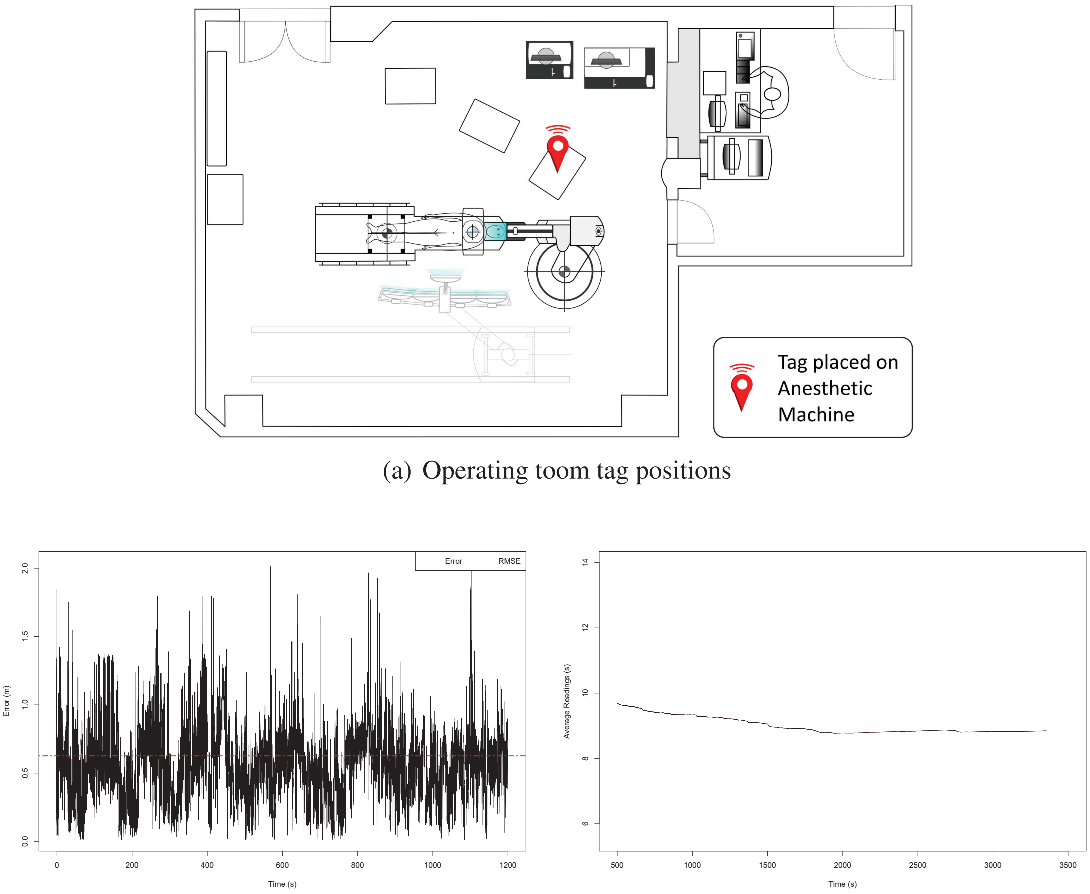
(d) Tag 3 errors

(e) Average readings per second

Figure 19: Scenario 1 experiment results

6.1.2 Accuracy Experiments: Scenario 2 – Emergency Procedure without using C-Arm

The second experiment was conducted during an emergency procedure in the OR. In this test, a single tag is placed in a safe position with authorization of the medical staff conducting the surgery. Figure 20(a) presents the blueprint of the room with the location of the tag, which was placed inside a compartment of an anesthetic machine, further, the figure presents the error estimations and the average readings per second. Since the anesthetic machine was moved during the procedure, as expected in an emergency surgery, the present error measurements were compromised, reporting an RMSE of 62 centimeters as observed in the noisy measures in Figure 20(b). The average readings per second on the other hand, were not affected as can be seen in Figure 20(c). Even in close proximity with the anesthetic machine functioning, staying in constant average 9 readings per second.



(a) Operating toom tag positions

(b) Tag errors

(c) Average readings per second

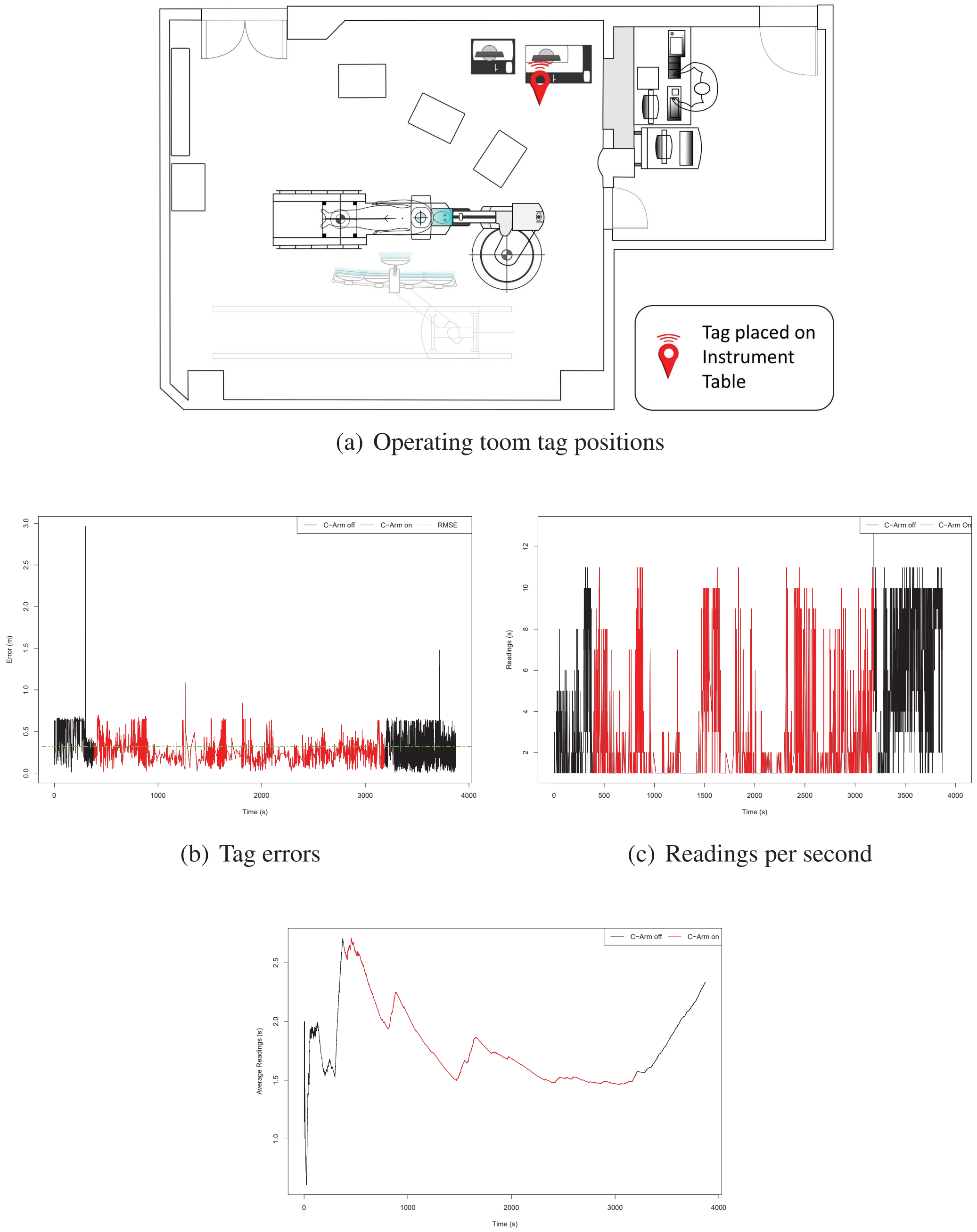
Figure 20: Scenario 2 experiment results

6.1.3 Accuracy Experiments: Scenario 3 – Angioplasty with C-Arm use

The third scenario consisted of monitoring a whole Angioplasty surgery with heavy C-Arm usage. Therefore, this experiment presents interesting insights on the impact of a hybrid surgical environment on UWB-based devices. The surgery lasted nearly 64 minutes and the C-Arm was active during the majority of the procedure.

In this procedure, the tag was placed on an instrument table, approximately two meters away from the C-Arm. The calculated RMSE of the tag was of 31 centimeters, the error measurements are presented on Figure 21(b). While larger than the errors observed in the first experiment, it still falls under the expected precision and will be reduced with the use of filters. Furthermore, the tag was located close to the room borders, where larger error measures are commonest. The most interesting data observed in this experiment is that, while the C-Arm is being used, the readings per second were affected and constantly dropped.

In Figure 21(a), the blueprint of the room with the tag position and three graphs featuring the observed errors, readings per second and average readings per second are depicted in Figures 21(c) and 21(d), respectively. During the C-Arm usage (presented as red lines in the three graphs), the frequency of readings is affected. The reasons of this interferences are to be studied during the future development of the project. Without the C-Arm usage, a mean of 4,8 readings per second is observed, reaching stable 6,1 readings per second right after the end of the C-Arm usage. However during the C-Arm use, an average of 1,37 readings per second is observed. In



(a) Operating toom tag positions

(b) Tag errors

(c) Readings per second

(d) Average readings per second

Figure 21: Scenario 3 experiment results

the average readings per second graph, it is noticeable that right after the end of the robotic arm usage, the readings rapidly started growing again.

6.2 Poses and RTLS Fusion Evaluation

This section analyses and discusses the experiments evaluating the proposed model for identifying poses using RTLS data. Four different experiments are discussed from this point onward. The first is a proof-of-concept of the proposed model, showing that is indeed feasible to use such data fusion for identifying poses. The three others analyse the proposed Ghosting heuristic for reducing the UWB position estimation error using estimated joints from the human poses, this analysis is conducted in a experimentation laboratory and in an hybrid operating room.

6.2.1 Pose-identification fusion proof-of-concept

The first experiment regards the pose-identification fusion proof-of-concept. This experiments' objective is to investigate the proposed idea of fusing the RTLS identification and position data with estimated human poses to track poses identity in a lightweight fashion and without vision-based markers. Since this experiments contain tag movement the observed error throughout the experiments is noticeably larger than in the RTLS evaluation. Furthermore, fewer error estimations are analyzed due to the low FPS which the CV module works on the computer used during the experiments.

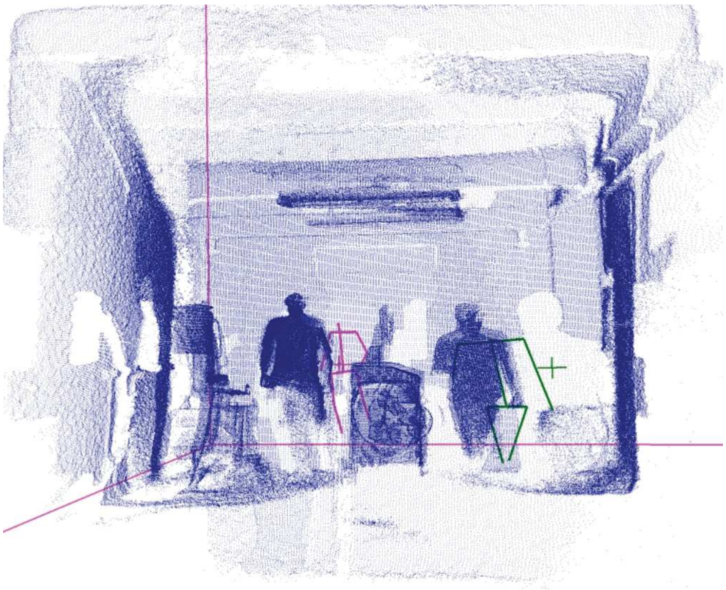


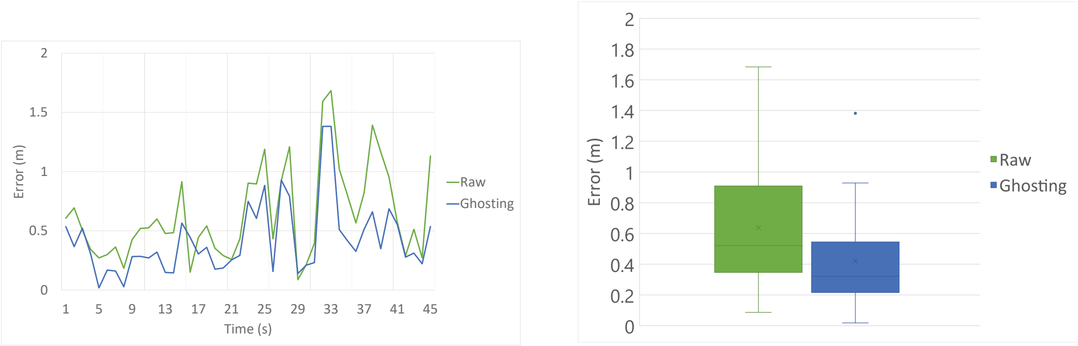
Figure 22: Viewer snapshot presenting fused HPE and RTLS data with two people. RTLS tags are drawn in the shape of crosses and poses are drawn on top of the respective person. A unique color is attributed to each identity. The coordinate axes are drawn as pink lines from the origin.

Figure 22 illustrates the fusion results. The image presents the estimated human poses, the depth data on the form of point clouds and the RTLS tags positions. Colored crosses depict the tag positions estimated by the RTLS, with colors used to differentiate individual tags. The people's poses are presented on the top of, or nearby, the respective person silhouette. Each detected pose has its color defined according to the identity of the RTLS tag attached to it. This proof-of-concept shows that it is indeed feasible to fuse pose and RTLS data to keep track of the pose identity.

Results indicate a considerable amount of identity switches, which occur due to errors on the estimation of RTLS tags positions. Occluded environments require further strategies to reduce the occurrence of identity switches. The proposed heuristics for reducing RTLS errors assess such problems. The next subsection discusses the results observed with this heuristic.

6.2.2 Ghosting heuristic evaluation

This subsection presents the results of the sensor fusion algorithm tests for error reduction detailed in the model, referred to as Ghosting. The experiments are conducted in two different environments, a experimentation lab and a real Operating Room (OR). The experiments in the experimentation laboratory are performed with one and two subjects while the OR experiment is solely performed with two. Next, observed results are discussed.



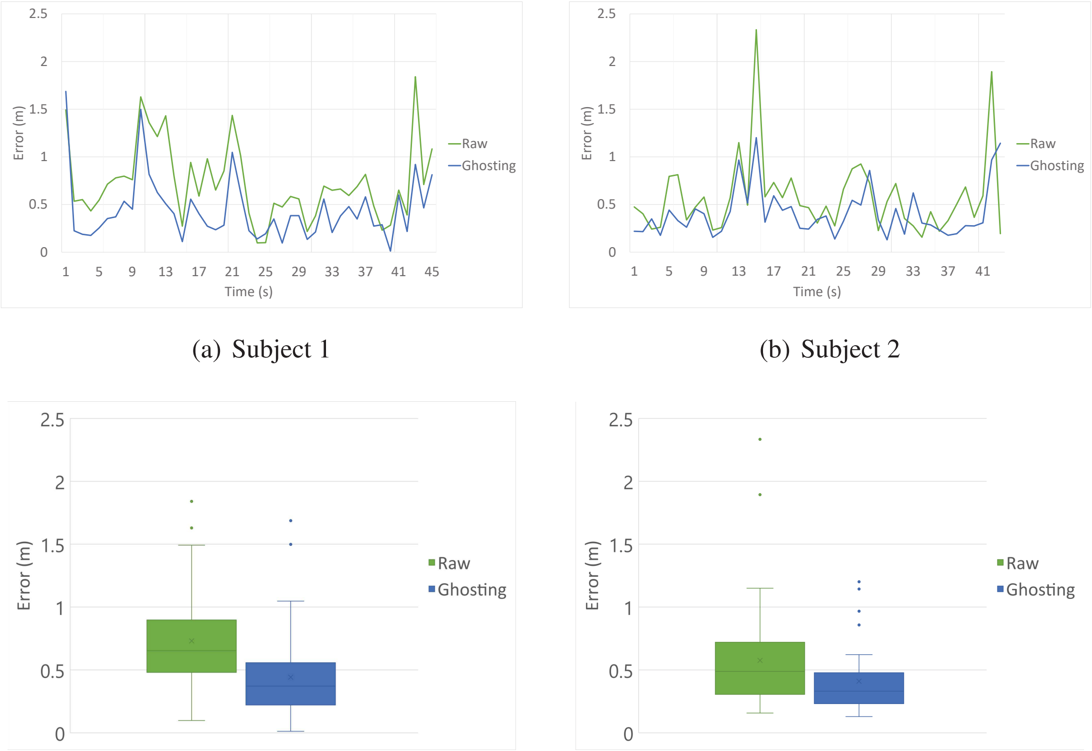
(a) Line chart comparing estimated errors by time.

(b) Box plot of error measurements.

Figure 23: Line chart and box plot of one subject experiment errors.

Figure 23 presents two different plots condensing processed data from experiments with one person. Throughout a 45 second tracking period, it is noticeable that filtered values regularly show errors below the raw estimated RTLS readings, as presented in Figure 23(a). And as shown in a box plot in Figure 23(b), nearly all data from the filtered interquartile range are below the median of the raw measurement. Further, despite one observed outlier error on filtered estimations, the maximum error observed on the filtered data is below one meter as shown by the upper whisker of the filtered data box plot, while raw measurements top nearly 1.8 meters of error.

Experiments with two subjects employ the same methodology as with one person. Figure 24 presents line charts and box plots depicting raw and filtered data of both subjects. On both line charts in Figures 24(a) and 24(b), a similar behaviour to the experiments observing a single



(a) Subject 1

(b) Subject 2

(c) Subject 1

(d) Subject 2

Figure 24: Line charts and box plots of raw and ghosting error measurements with two subjects.

person is noticed. Filtered values continuously outperform raw RTLS readings, as the filtered tag values are drawn closer to the correct poses joints by using the last frames joints positions as pre-states. While identity switches analysis is not the focus of the current work, the 41-second mark depicts a potential switch since both raw and filtered estimations present abnormal behavior. Despite this, the majority of error measurements are still in conformity with previous results with one subject. As seen in the box plots in Figures 24(c) and 24(d), of Subject 1 and 2 respectively, a larger amount of outliers were observed due to possible identity switches. However, once again the same behavior is observed accordingly to the previous experiment, interquartile filtered errors remain below the median of raw measurements, for both subjects. Also, the maximum filtered error is slightly above one meter, nearly 50 centimeters below the maximum raw error for Subject 1. For Subject 2, raw measurements were more precise, resulting in more condensed results, with smaller interquartile boxes and whiskers. Even though results indicate lower raw error measurements, filtering further reduced the errors on the same fashion to both previous analysis of Person 1 data.

Lastly, the Root Mean Square Error (RMSE) of all error estimates from these two data fusion experiments are presented. Figure 25 depicts the RMSE of both experiments. For the Subject 1 experiment, an error reduction of 46.15% is obtained, reducing the raw measurements error from 39.05 to 21.75 centimeters with the proposed heuristics. Since the RMSE is a metric sensible to outliers, as expected, results indicate a noticeable increase in the error for the ex-

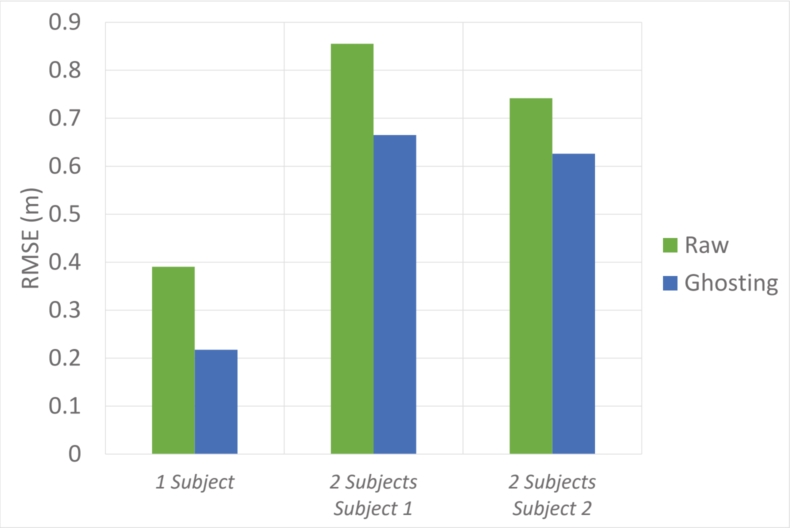


Figure 25: Random Mean Square Error on both experiments.

periment with two subjects. For Subject 1 and 2, an error reduction of 22.24% and 15.57% is observed, reducing the errors from 85.51 to 66.49 centimeters for Subject 1 and from 74.16 to 62.61 centimeters for Subject 2.

Regarding the experiments conducted in the hospital. As expected due to the large amount of equipment and metallic surfaces, larger errors on the tag estimations during movement are observed. And as expected accordingly to the experiments conducted in the laboratory, the ghosting heuristic once again achieved relevant results which are discussed next. Figure 26 presents a color point cloud of one of the frames obtained during the experiment.

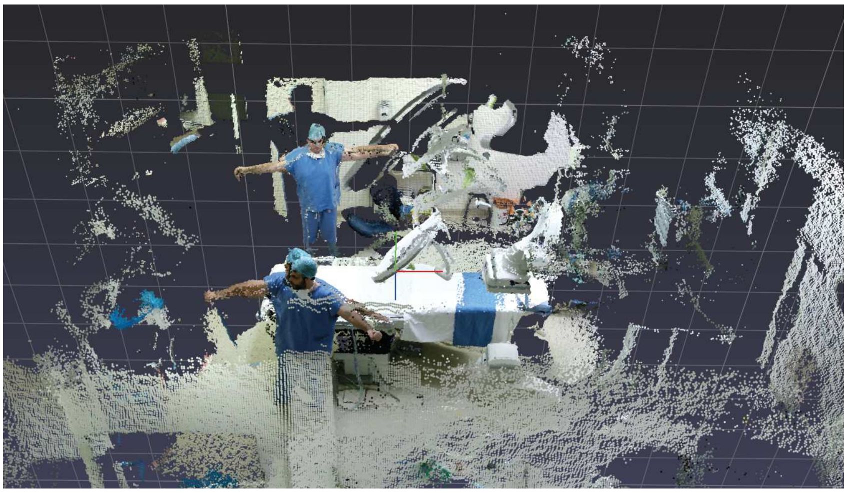
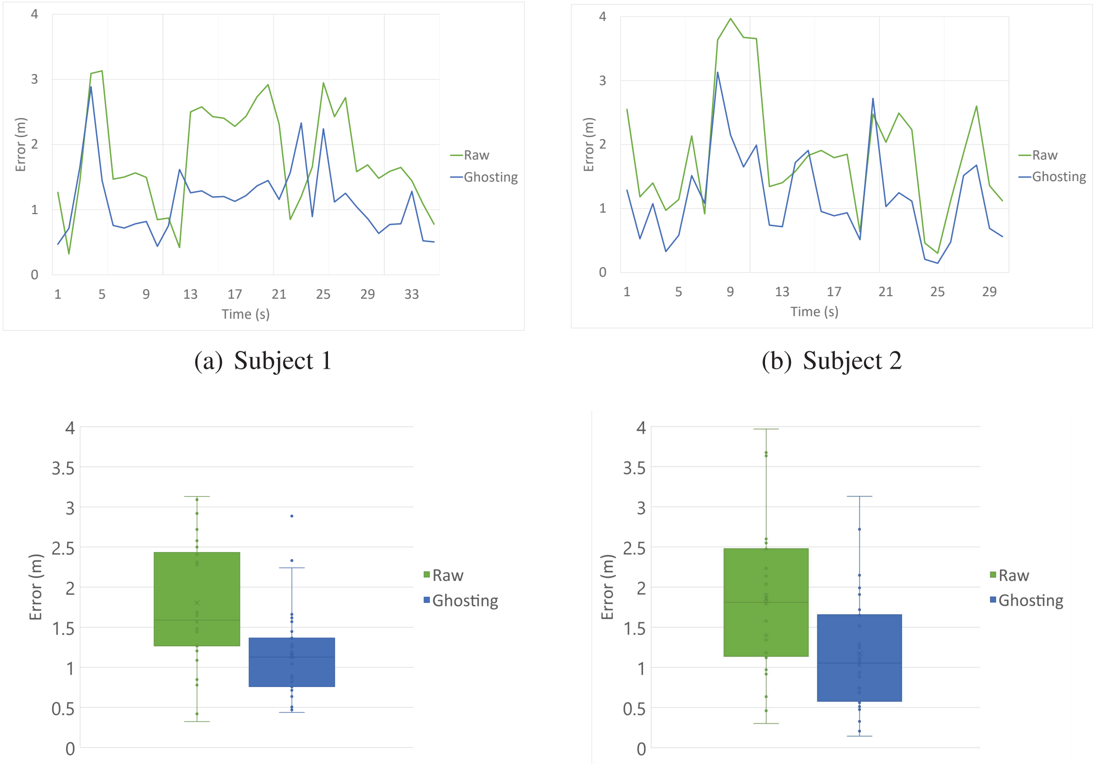


Figure 26: Color point cloud obtained during hospital experiments.

The tests conducted in the OR follow the same methodology as the ones in the experimentation scenario. All of the parameters are the same except for the cameras placement and room size. On Figure 27 the processed data from the experiments is presented in the same fashion as the previous ones. Figures 27(a) and 27(b) show that the ghosting filter constantly achieves



(a) Subject 1

(b) Subject 2

(c) Subject 1

(d) Subject 2

Figure 27: Line charts and box plots of raw and ghosting error measurements with two subjects.

lower error measurements than the raw readings for both subjects. As previously stated, larger error measurements can be observed and are more clearly detailed in Figures 27(c) and 27(d) which depict the box plots for subject 1 and subject 2. Once again all ghosting filter interquartile values are below the median of the raw measurements. Showing the consistency of the proposed technique.

Lastly, Figure 28 presents the RMSE of these results. The RMSE of the raw errors are of 1,34 and 1,36 meters for subject 1 and 2, respectively. While the RMSE of the ghosting errors are of 1,07 meters for subject 1 and 1,08 meters for subject 2. This corresponds to a reduction of 21,15% and 21,59%. In this setup the observed RMSE achieved similar results for both subjects, and the consistency of the proposed filter is once again reinforced.

In this chapter, the results of the experiments assessing the proposed model are detailed. Experiments regarding RF and pose data fusion present encouraging results. These results can be enhanced with better hardware for processing data at higher FPS and with domain-specific trained CNN for body part detecting. Even with low FPS count as achieved with the used hardware, and using a generic dataset-trained CNN for the HPE, results of more than 45% of position RTLS estimation error reduction are achieved. The Ghosting heuristic achieves larger error reduction than that of other works that focus on reducing position error estimations by fusing RF and CV data, such as from (STURARI et al., 2016), which already had obtained sensible error reductions. The authors report an average obtained error of 70 centimeters, while

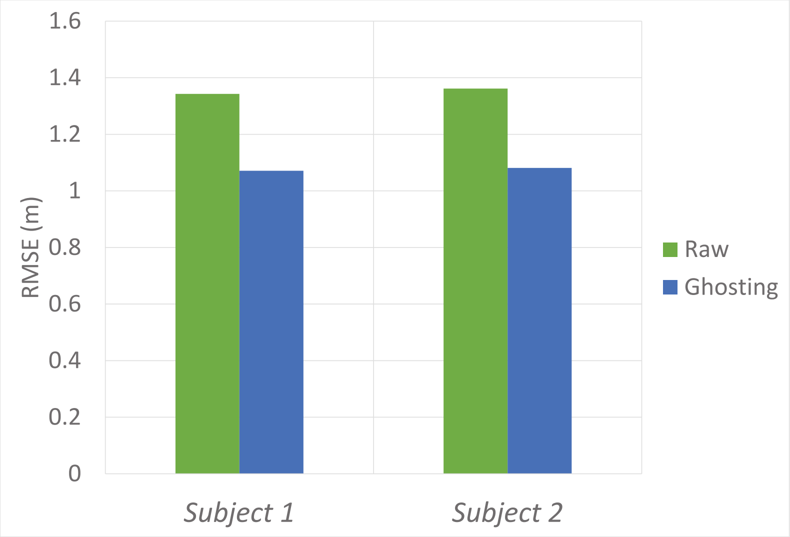


Figure 28: Random Mean Square Error of ghosting and unfiltered errors at the hospital experiment.

in this proposal of fusing estimated body parts and RF data, the best case returned an average of 21.75 centimeters of error.

7 CONCLUSION

While the capacity of human pose estimators increased in the last few years, keeping track of the estimated poses' identities is still a considerable challenge in the HPE area, due to occlusions and constant in and out-of-view tracked subjects. Such problems arise due to the principle of camera devices, which can only be assessed by either deploying more equipment or applying techniques based on image processing and temporal consistency. RTLS on the other hand, contains solutions to this problem, permitting constant monitoring of tracked subjects identity and location, not suffering from occlusion and covering larger areas than cameras. While not as millimetric accurate as state-of-the-art RGBD devices, point level accuracy is already achievable with market-available RTLS. This fact not only enables the fusion of data for identifying poses, but also the reduction of the estimation errors from RTLS through data fusion. Since recognizing activities is a highly contextual problem, keeping track of information such as pose identities is crucial. In this context, this paper presents the Baptizo model which attacks the specific problem of identifying poses. The model consists of fusing poses and RTLS data for keeping track of the estimated human poses identity, leading to a win-win situation. The main scientific contributions of this thesis are two: the technique for identifying human poses through sensor fusion, and the proposed Ghosting heuristic for reducing errors on the position estimation of radiofrequency devices also through data fusion.

The proposed model is composed of independent modules that separately consume and process the RF and RGBD sensors produced data. The core of the proposed model lives in the sensor fusion module, which consumes the previous modules processed pose and position data and fuses them, into a single data frame, an identified human pose data. Each module can be deployed on the same or separated nodes, depending on infrastructural needs. A proof-of-concept and other experiments are conducted in a constrained and real scenario to test the feasibility of the proposed model. With this proposed model, recovering from occlusions and out-of-view tracked subjects is made simpler. While keeping track of identities would demand vast amounts of computational resources by applying CV-based strategies, the proposed approach is lightweight.

Moreover, experiments show that the Ghosting heuristic of fusing RF position data and RGBD estimated poses is not only feasible, but promising. Errors were reduced by 46%, significantly increasing the reliability of the final fused data. In comparison to other state-of-the-art works fusing data from radiofrequency and camera devices to reduce estimation errors, we obtain an additional reduction of 31%. Furthermore, as aforementioned, there are means to further the error reduction, which are to be assessed in future works together with identity switches tracking.

This current work proposes a novel way to extract identified pose data in a lightweight id tracking fashion, enabling the extraction of valuable data for carrying out activity recognition strategies. While the proposed method still is mildly-intrusive, demanding that people carry the

RTLS tag attached or close to a given body part, it is less error-prone and less intrusive than other marker-based video tracking solutions. The proposed work enables further research and development of applications on several areas, such as gaming, surveillance and healthcare. This way, paving the way for improved entertainment, security and quality of life.

REFERENCES

ABADI, M.; BARHAM, P.; CHEN, J.; CHEN, Z.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; IRVING, G.; ISARD, M. et al. Tensorflow: a system for large-scale machine learning. In: OSDI, 2016. **Anais...** [S.l.: s.n.], 2016. v. 16, p. 265–283.

AGGARWAL, J. K.; XIA, L. Human activity recognition from 3d data: a review. **Pattern Recognition Letters**, [S.l.], v. 48, p. 70–80, 2014.

AKYILDIZ, I. F.; SU, W.; SANKARASUBRAMANIAM, Y.; CAYIRCI, E. Wireless sensor networks: a survey. **Computer networks**, [S.l.], v. 38, n. 4, p. 393–422, 2002.

ANTUNES, R. S.; SEEWALD, L. S.; RODRIGUES, V. F.; COSTA, C. A.; GONZAGA JR., L.; RIGHI, R. R.; MAIER, A.; ESKOFIER, B.; OLLENSCHLäGER, M.; NADERI, F.; FAHRIG, R.; BAUER, S.; KLEIN, S.; CAMPANATTI JR., G. A Survey of Sensors in Healthcare Workflow Monitoring. **ACM Comput. Surv.**, New York, NY, USA, v. 51, n. 2, 2018.

ATZORI, L.; IERA, A.; MORABITO, G. The internet of things: a survey. **Computer networks**, [S.l.], v. 54, n. 15, p. 2787–2805, 2010.

BAUER, S.; SEITEL, A.; HOFMANN, H.; BLUM, T.; WASZA, J.; BALDA, M.; MEINZER, H.-P.; NAVAB, N.; HORNEGGER, J.; MAIER-HEIN, L. Real-time range imaging in health care: a survey. In: **Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications**. [S.l.]: Springer, 2013. p. 228–254.

BOULOS, M. N. K.; BERRY, G. Real-time locating systems (RTLS) in healthcare: a condensed primer. **International journal of health geographics**, [S.l.], v. 11, n. 1, p. 25, 2012.

CAO, Z.; SIMON, T.; WEI, S.-E.; SHEIKH, Y. Realtime multi-person 2d pose estimation using part affinity fields. **arXiv preprint arXiv:1611.08050**, [S.l.], 2016.

CHEN, M.; GONZALEZ, S.; VASILAKOS, A.; CAO, H.; LEUNG, V. C. Body area networks: a survey. **Mobile networks and applications**, [S.l.], v. 16, n. 2, p. 171–193, 2011.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: CONFERENCE ON COMPUTER VISION & PATTERN RECOGNITION (CVPR'05), 2005. **Anais...** [S.l.: s.n.], 2005. v. 1, p. 886–893.

DARDARI, D.; CLOSAS, P.; DJURIC, P. M. Indoor Tracking: theory, methods, and technologies. **IEEE Trans. Vehicular Technology**, [S.l.], v. 64, n. 4, p. 1263–1278, 2015.

DRAZOVICH, R. Sensor fusion in tactical warfare. In: COMPUTERS IN AEROSPACE CONFERENCE, 4., 1983. **Anais...** [S.l.: s.n.], 1983. p. 2398.

FARID, Z.; NORDIN, R.; ISMAIL, M. Recent advances in wireless indoor localization techniques and system. **Journal of Computer Networks and Communications**, [S.l.], v. 2013, 2013.

FLEURET, F.; BERCLAZ, J.; LENGAGNE, R.; FUA, P. Multicamera people tracking with a probabilistic occupancy map. **IEEE transactions on pattern analysis and machine intelligence**, [S.l.], v. 30, n. 2, p. 267–282, 2008.

FORTINO, G.; GHASEMZADEH, H.; GRAVINA, R.; LIU, P. X.; POON, C. C.; WANG, Z. Advances in multi-sensor fusion for body sensor networks: algorithms, architectures, and applications. **Information Fusion**, [S.l.], v. 45, p. 150–152, 2019.

GIRSHICK, R.; SHOTTON, J.; KOHLI, P.; CRIMINISI, A.; FITZGIBBON, A. Efficient regression of general-activity human poses from depth images. In: COMPUTER VISION (ICCV), 2011 IEEE INTERNATIONAL CONFERENCE ON, 2011. **Anais...** [S.l.: s.n.], 2011. p. 415–422.

GRAVINA, R.; ALINIA, P.; GHASEMZADEH, H.; FORTINO, G. Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges. **Information Fusion**, [S.l.], v. 35, p. 68–80, 2017.

HARTMANN, F.; SCHLAEFER, A. Feasibility of touch-less control of operating room lights. **International journal of computer assisted radiology and surgery**, [S.l.], v. 8, n. 2, p. 259–268, 2013.

HAUTE, T.; POORTER, E.; CROMBEZ, P.; LEMIC, F.; HANDZISKI, V.; WIRSTRÖM, N.; WOLISZ, A.; VOIGT, T.; MOERMAN, I. Performance analysis of multiple Indoor Positioning Systems in a healthcare environment. **International journal of health geographics**, [S.l.], v. 15, n. 1, p. 7, 2016.

HELLMICH, T. R.; CLEMENTS, C. M.; EL-SHERIF, N.; PASUPATHY, K. S.; NESTLER, D. M.; BOGGUST, A.; ERNSTE, V. K.; MARISAMY, G.; KOENIG, K. R.; HALLBECK, M. S. Contact tracing with a real-time location system: a case study of increasing relative effectiveness in an emergency department. **American journal of infection control**, [S.l.], v. 45, n. 12, p. 1308–1311, 2017.

HIGHTOWER, J.; BORRIELLO, G. Location systems for ubiquitous computing. **Computer**, [S.l.], v. 34, n. 8, p. 57–66, 2001.

HOLTE, M. B.; TRAN, C.; TRIVEDI, M. M.; MOESLUND, T. B. Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. **IEEE Journal of selected topics in signal processing**, [S.l.], v. 6, n. 5, p. 538–552, 2012.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. Mobilenets: efficient convolutional neural networks for mobile vision applications. **arXiv preprint arXiv:1704.04861**, [S.l.], 2017.

HUSSMANN, S.; RINGBECK, T.; HAGEBEUKER, B. A performance review of 3D TOF vision systems in comparison to stereo vision systems. In: **Stereo vision**. [S.l.]: InTech, 2008.

INSAFUTDINOV, E.; PISHCHULIN, L.; ANDRES, B.; ANDRILUKA, M.; SCHIELE, B. Deepercut: a deeper, stronger, and faster multi-person pose estimation model. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2016. **Anais...** [S.l.: s.n.], 2016. p. 34–50.

IQBAL, U.; MILAN, A.; GALL, J. Posetrack: joint multi-person pose estimation and tracking. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2017. **Proceedings...** [S.l.: s.n.], 2017. p. 2011–2020.

KANG, H.; HOSTETLER, S.; DEVAPRIYA, P.; BANCIU, M.; ANDREWS, Z. RTLS and EHR Enabled Workflow Modeling in the Emergency Department. In: IIE ANNUAL CONFERENCE. PROCEEDINGS, 2014. **Anais...** [S.l.: s.n.], 2014. p. 3112.

KAUTZ, T.; GROH, B. H.; ESKOFIER, B. M. Augmented motion models for constrained position tracking with Kalman filters. In: INFORMATION FUSION (FUSION), 2016 19TH INTERNATIONAL CONFERENCE ON, 2016. **Anais...** [S.l.: s.n.], 2016. p. 849–854.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2012. **Anais...** [S.l.: s.n.], 2012. p. 1097–1105.

LACHAT, E.; MACHER, H.; MITTET, M.; LANDES, T.; GRUSSENMEYER, P. First experiences with Kinect v2 sensor for close range 3D modelling. **The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, [S.l.], v. 40, n. 5, p. 93, 2015.

LAHTELA, A.; HASSINEN, M.; JYLHA, V. RFID and NFC in healthcare: safety of hospitals medication care. In: PERVASIVE COMPUTING TECHNOLOGIES FOR HEALTHCARE, 2008. PERVASIVEHEALTH 2008. SECOND INTERNATIONAL CONFERENCE ON, 2008. **Anais...** [S.l.: s.n.], 2008. p. 241–244.

LATRÉ, B.; BRAEM, B.; MOERMAN, I.; BLONDIA, C.; DEMEESTER, P. A survey on wireless body area networks. **Wireless Networks**, [S.l.], v. 17, n. 1, p. 1–18, 2011.

LEE, J.; BAGHERI, B.; KAO, H.-A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. **Manufacturing letters**, [S.l.], v. 3, p. 18–23, 2015.

LI, X.; TENG, J.; ZHAI, Q.; ZHU, J.; XUAN, D.; ZHENG, Y. F.; ZHAO, W. Ev-human: human localization via visual estimation of body electronic interference. In: INFOCOM, 2013 PROCEEDINGS IEEE, 2013. **Anais...** [S.l.: s.n.], 2013. p. 500–504.

LICIOTTI, D.; CONTIGIANI, M.; FRONTONI, E.; MANCINI, A.; ZINGARETTI, P.; PLACIDI, V. Shopper analytics: a customer activity recognition system using a distributed rgb-d camera network. In: INTERNATIONAL WORKSHOP ON VIDEO ANALYTICS FOR AUDIENCE MEASUREMENT IN RETAIL AND DIGITAL SIGNAGE, 2014. **Anais...** [S.l.: s.n.], 2014. p. 146–157.

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft coco: common objects in context. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2014. **Anais...** [S.l.: s.n.], 2014. p. 740–755.

LIU, H.; DARABI, H.; BANERJEE, P.; LIU, J. Survey of wireless indoor positioning techniques and systems. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, [S.l.], v. 37, n. 6, p. 1067–1080, 2007.

MANDELJC, R.; KOVAČIČ, S.; KRISTAN, M.; PERŠ, J. et al. Tracking by identification using computer vision and radio. **Sensors**, [S.l.], v. 13, n. 1, p. 241–273, 2012.

MCAFEE, A.; BRYNJOLFSSON, E.; DAVENPORT, T. H.; PATIL, D.; BARTON, D. Big data: the management revolution. **Harvard business review**, [S.l.], v. 90, n. 10, p. 60–68, 2012.

NEWELL, A.; YANG, K.; DENG, J. Stacked hourglass networks for human pose estimation. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2016. **Anais...** [S.l.: s.n.], 2016. p. 483–499.

PAPAIOANNOU, S.; MARKHAM, A.; TRIGONI, N. Tracking people in highly dynamic industrial environments. **IEEE Transactions on Mobile Computing**, [S.l.], v. 16, n. 8, p. 2351–2365, 2017.

PAPAIOANNOU, S.; WEN, H.; MARKHAM, A.; TRIGONI, N. Fusion of radio and camera sensor data for accurate indoor positioning. In: MOBILE AD HOC AND SENSOR SYSTEMS (MASS), 2014 IEEE 11TH INTERNATIONAL CONFERENCE ON, 2014. **Anais...** [S.l.: s.n.], 2014. p. 109–117.

PAU, L.-F. Sensor data fusion. **Journal of Intelligent and Robotic Systems**, [S.l.], v. 1, n. 2, p. 103–116, 1988.

PERŠ, J.; KRISTAN, M.; KOVAČIČ, S. et al. Fusion of non-visual modalities into the Probabilistic Occupancy Map framework for person localization. In: DISTRIBUTED SMART CAMERAS (ICDSC), 2011 FIFTH ACM/IEEE INTERNATIONAL CONFERENCE ON, 2011. **Anais...** [S.l.: s.n.], 2011. p. 1–6.

RAWAT, P.; SINGH, K. D.; CHAOUCHI, H.; BONNIN, J. M. Wireless sensor networks: a survey on recent developments and potential synergies. **The Journal of supercomputing**, [S.l.], v. 68, n. 1, p. 1–48, 2014.

SCHMALZ, C.; FORSTER, F.; SCHICK, A.; ANGELOPOULOU, E. An endoscopic 3D scanner based on structured light. **Medical image analysis**, [S.l.], v. 16, n. 5, p. 1063–1072, 2012.

SEEWALD, L. A.; RODRIGUES, V. F.; OLLENSCHLÄGER, M.; ANTUNES, R. S.; COSTA, C. A. da; ROSA RIGHI, R. da; SILVEIRA JR, L. G. da; MAIER, A.; ESKOFIER, B.; FAHRIG, R. Toward analyzing mutual interference on infrared-enabled depth cameras. **Computer Vision and Image Understanding**, [S.l.], 2018.

SHIREHJINI, A. A. N.; YASSINE, A.; SHIRMOHAMMADI, S. Equipment location in hospitals using RFID-based positioning system. **IEEE Transactions on information technology in biomedicine**, [S.l.], v. 16, n. 6, p. 1058–1069, 2012.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, [S.l.], 2014.

STAUFFER, C.; GRIMSON, W. E. L. Adaptive background mixture models for real-time tracking. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CAT. NO PR00149), 1999., 1999. **Proceedings...** [S.l.: s.n.], 1999. v. 2, p. 246–252.

STURARI, M.; LICIOTTI, D.; PIERDICCA, R.; FRONTONI, E.; MANCINI, A.; CONTIGIANI, M.; ZINGARETTI, P. Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. **Pattern Recognition Letters**, [S.l.], v. 81, p. 30–40, 2016.

TENG, J.; ZHANG, B.; ZHU, J.; LI, X.; XUAN, D.; ZHENG, Y. F. EV-Loc: integrating electronic and visual signals for accurate localization. **IEEE/ACM Transactions on Networking (TON)**, [S.l.], v. 22, n. 4, p. 1285–1296, 2014.

WASENMÜLLER, O.; STRICKER, D. Comparison of kinect v1 and v2 depth images in terms of accuracy and precision. In: ASIAN CONFERENCE ON COMPUTER VISION, 2016. **Anais. . .** [S.l.: s.n.], 2016. p. 34–45.

WEISER, M. The Computer for the 21 st Century. **Scientific american**, [S.l.], v. 265, n. 3, p. 94–105, 1991.

WU, X.; ZHU, X.; WU, G.-Q.; DING, W. Data mining with big data. **IEEE transactions on knowledge and data engineering**, [S.l.], v. 26, n. 1, p. 97–107, 2013.

XIU, Y.; LI, J.; WANG, H.; FANG, Y.; LU, C. Pose Flow: efficient online pose tracking. **arXiv preprint arXiv:1802.00977**, [S.l.], 2018.

YANG, L.; ZHANG, L.; DONG, H.; ALELAIWI, A.; EL SADDIK, A. Evaluating and improving the depth accuracy of Kinect for Windows v2. **IEEE Sensors Journal**, [S.l.], v. 15, n. 8, p. 4275–4285, 2015.

ZHAO, Y.; LIU, Y.; NI, L. M. VIRE: active rfid-based localization using virtual reference elimination. In: PARALLEL PROCESSING, 2007. ICPP 2007. INTERNATIONAL CONFERENCE ON, 2007. **Anais. . .** [S.l.: s.n.], 2007. p. 56–56.