TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Konferencia UAV 2023

Kód projektu: 313011V422
„Inteligentné operačné a spracovateľské systémy pre UAV“

ISBN 978-80-553-4389-1

2023

Zborník z konferencie UAV 2023

Vydavateľ: Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9, 040 01 Košice, Slovenská republika

Dátum vydania: máj 2023

Tlač: 50 kusov CD

Jazyk: slovenský, anglický

Strán: 86

Editor: Ing. Miroslav Michalko, PhD.

ISBN: 978-80-553-4389-1

O konferencii UAV 2023

Projektová konferencia UAV 2023 bola organizovaná na pôde Technickej univerzity v Košiciach, v rámci prezentačných priestorov Univerzitného vedeckého parku TECHNICOM, v dňoch 11. – 12. 5. 2023, v rámci projektu:

Názov projektu:	Inteligentné operačné a spracovateľské systémy pre UAV
Kód projektu:	313011V422
Prijímateľ NFP:	GLOBESY, s.r.o.
Partneri:	Qintec a.s.
	YMS, a.s.
	Technická univerzita v Košiciach (TUKE)
	Žilinská univerzita v Žiline (UNIZA)

(ďalej len „projekt UAV").

Projekt UAV je spolufinancovaný z prostriedkov Európskeho fondu regionálneho rozvoja v rámci Operačného programu Integrovaná infraštruktúra.

EURÓPSKA ÚNIA
Európske štrukturálne a investičné fondy
OP Integrovaná infraštruktúra 2014 – 2020

Riadiaci orgán:

MINISTERSTVO DOPRAVY SLOVENSKEJ REPUBLIKY

V zastúpení na základe splnomocnenia:

VÝSKUMNÁ AGENTÚRA

Obsah

Image data annotated by objects distances

[1,2]*Peter Vojtas*

[1] Globesy Ltd, Framborska 58, Zilina, Slovakia
[2] Dpt. Software Engineering, Charles University, Prague, Czechia

pvojtas@gmail.com

Abstract **— The main motivation of this paper is to study technologies for the protection of healthcare personnel on the front line and the operation of medical facilities during the spreading of the disease Covid-19. Covid is over, nevertheless, some technical problems remained. We reduce here ourselves to an estimation of the distance between objects recognized by fixed cameras or UWB sensors in an indoor environment (not using GPS or a smartphone). The main problem which crystallized was the need for a verified distance annotated data set. The main idea is to use synchronized data from video and localization sensor signals in a controlled experiment. The hope is that such an annotated data set can be used for training a tool able to recognize distances and positions. Yet another problem arose - transferability to new environments. To support this, we suggest using additional knowledge about the environment and changing the installation phase. This is a work-in-progress paper.**

Keywords **— synchronizing UWB and video data, controlled experiments, indoor localization, environment knowledge, deep neural learning,**

I. Introduction, motivation

The main motivation of this paper is to study technologies for the protection of healthcare personnel on the front line and the operation of medical facilities during the spreading of the disease Covid-19. Our main interest here is indoor technologies for contact tracing and distance estimation.

According to [5], some of the main challenges experienced by the different organizations during the COVID-19 pandemic were the lack of preparedness of several healthcare and information technology companies and institutions. In the edited volume [5] contributors discuss both the positive and negative impacts of the COVID-19 pandemic on the healthcare and technology sectors. Contact tracing based on smartphone application is mentioned in Chapter 13. A much more detailed discussion on smartphone tracing and distance estimation is in [6]. These have a smaller impact in an indoor environment. Quite good indoor results are referred to using BLE based AoA (angle of arrival) method[1].

We reduce here this to estimating the distance between objects recognized by fixed cameras and/or sensors in an indoor environment. The main problem which crystallized was the need for a distance annotated data set. Annotation and labeling of raw data — images and videos — for machine learning (ML) models are the most time-consuming and laborious, albeit essential, phase of any computer vision project[2]. As further mentioned by [8], this is mainly restricted to labeling by categories. Successful annotation outcomes ensure an ML model can 'learn' from these training data, solving the problems organizations and ML team leaders set out to solve. Nevertheless, we were not able to find image and/or video data annotated also by distances between recognized objects.

In this work-in-progress paper, we describe initial considerations towards indoor distance and/or position annotation of objects (typically humans) in videos. We study several methods of indoor distance and/or position annotation of objects. The hope is that by having such annotated data we will be able to update a neural network already trained e.g. for object detection/recognition (see [3]).

Yet another problem arose - transferability to new environments. To support this, we suggest using additional knowledge about the environment and changing the installation phase ([2, 7]).

[1] https://www.bluetooth.com/learn-about-bluetooth/feature-enhancements/direction-finding/
[2] https://encord.com/blog/guide-to-outsourcing-data-labeling-for-machine-learning/

We distinguish 3 types of indoor application domains. First are those equipped with both cameras and positioning sensors. Second have only cameras. Third do not allow the use of cameras and allow only positioning sensors (nevertheless, during installation it can be useful a restricted use of cameras).

There are three basic use-cases:

The first use-case requires to have an automated system for indoor (without GPS and smartphone) estimation of the mutual distance between moving objects from a fixed camera. A typical example is safety regulation during corona pandemic requesting to keep some minimal distance between persons.

The second use-case is a system that tries to prevent users to get too close to some dangerous area (e.g. the flowing iron in a steel factory, a highly polluted area, etc.). Another example is an ICU-intensive care unit (intensive therapy unit or intensive treatment unit or critical care unit) i.e. a special department of a hospital or health care facility that provides intensive care medicine - that provides treatment and monitoring for people who are in serious health conditions. Here a patient may behave dangerously.

The third use-case is the detection of intruders or persons without permission (e.g. without a sensor).

The general idea is that one can train a system for event detection. We decided to go in the direction to base such systems on distance estimation (either mutual distance between persons or distance to a dangerous area or moving away from a safe area (e.g. a bed in ICU)).

The basic idea is to have a system that detects/recognizes humans in a video (e.g. with a bounding box). We will design several methods of distance estimations. In a controlled experiment environment, we will evaluate and compare these methods. We would like to provide experiments in different environments to avoid over-training and guarantee transferability.

We distinguish two groups of methods. First relies only on the video signal and maybe some additional knowledge. The second uses positioning sensors. Sometimes a combination of both can be possible.

In the application, we distinguish between installation and deployment. At the time of installation in a new environment, we can need also the usage of videos even in domains where videos are forbidden. Another case can be the situation where participants cannot be equipped with sensors, nevertheless in organized experiments, we can work with participants wearing sensors.

Original motivation comes from the funding project. Namely, the task is to:

- Develop tool prototypes for the use of sensors and cameras in the monitoring of medical facilities and key facilities during the pandemic and the detection of potentially infected persons (disease-tracking) in these areas
- Develop a prototype for monitoring personnel and material using machine learning

There is yet another task of the project. Namely, to develop a machine learning-based tool prototype to support the planning of front-line medical personnel and to ensure the operation of medical facilities during a pandemic.

Having the above-mentioned task, and hoping to successfully finish the work started by this work in progress, we can hope to contribute to the development of prototype tools to support the planning of front-line medical personnel and to ensure the operation of medical facilities during a pandemic. Nevertheless, so far, this is out of the scope of this paper.

The main contribution will be a video data set annotated with distances. Several methods for estimating distances in a new environment together with a model of knowledge about this environment.

This is a work-in-progress paper. We do not deal here with using these annotated data to train the distance estimation system. Our task here is only a proposal of methods limited only to the creation of distance annotated data and evaluation and comparison of several methods concerning precision. We base our proposal on some initial experiments with BLE and UWB sensors. Our objects are assumed to move slowly and precision has to be useful for the respective use-case.

The paper is organized as follows: In Chapter 2 we mention our previous work on object detection/recognition without additional labeling for a new environment. Chapter 3 is describing some baseline optical methods for distance estimation and the organization of a controlled experiment. Chapter 4 offers possible requirements for additional environment knowledge to enhance the transferability of our tool. Chapter 5 gives a short overview of UWB sensors we acquired and experimented (in this initial phase we do not mention any results on these). Chapter 6 concludes this work-in-progress paper.

II. A METHOD FOR OBJECT DETECTION/RECOGNITION WITHOUT ADDITIONAL LABELING FOR A NEW ENVIRONMENT

For object detection/recognition we use the results of [1]. The main goal of [1] was to create a fast emergency aid system for object detection in SME (small medium enterprise) industrial premises from CCTV-IP cameras. We used a variety of deep neural networks pre-trained in a smart city environment without any annotation, training, or human intervention from an industrial environment. We introduced a pseudo-ground truth method to decide object detection. A significant number of pre-trained computer learning models from the computer vision category are freely available on the Internet. Our primary interest is the area of industry. However, we did not find any training set from the industrial domain, e.g., office, production line, warehouse, parking lot, etc. Obtaining annotated data needed for training is a time-consuming activity that requires user intervention, and therefore, we want to avoid it. We chose several pre-trained models based on the COCO reference dataset [4]. This dataset contains more than 330,000 images containing annotations of 1.5 million instances of objects from 80 classes. But we can say that it is more from the area of "smart cities" and none of our target areas. Of course, we are aware that the accuracy of such models is probably significantly worse than in the case of their training in a specific environment (warehouse, office, etc.).

In [1] we considered a situation where we do not have any human-annotated training data for object detection in an industrial environment without any human intervention. This has led us to a concept of “pseudo ground truth”. Pseudo ground truth PGT_3 is created by a heuristic process considering a correct object detection to be the one where at least three models agreed (lower index 3 in PGT_3 refers to the number of models required to agree on an instance), see also [9].

III. SOME INITIAL EXPERIMENTS – THE BASELINE

Here we describe some baseline optical methods for distance estimation and the organization of a controlled experiment.

In Figure 1 we see the hallway where the initial experiments were carried out.

The main idea was to have a controlled experiment to be able to verify objects’ distances by several methods.

To achieve this we used ground tags in a regular 1m distance (see G_1, G_2, G_3, G_4…). The image is only illustrative and does not fully describe all aspects of controlling the experiment. It shows two persons P_1 and P_2.

We start a video. The scenario assumes persons are slowly moving along ground tags and hence we know where in which time they stayed. In the initial phase, persons halt on the ground tag a make position recognition easier. In processing video, we use a deep neural network to recognize persons (denoted by a bounding box). Without loss of generality, we assume that the bottom side of the box denotes the person’s position along ground tags.

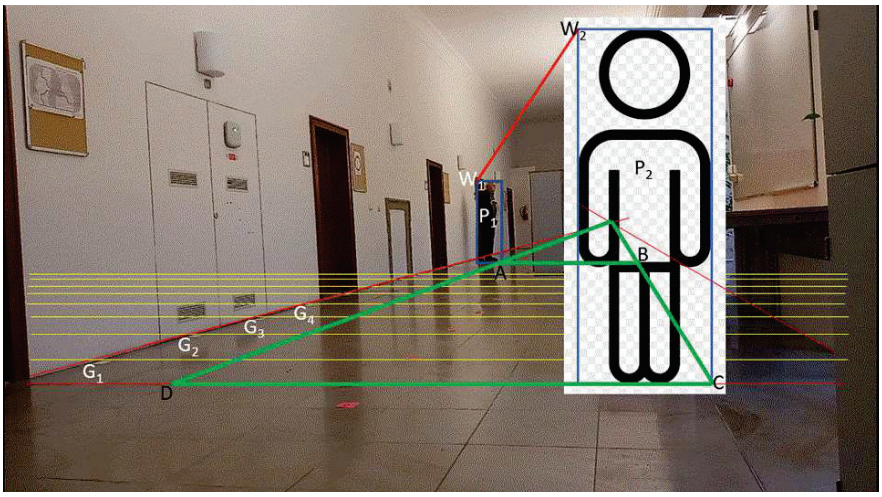

Fig. 1 An illustrative image of a calculation of distance between persons based on pixel distance weight and a rectangle in perspective and ground tags coordinates

The first idea for the calculation of distance between people is very straightforward. Assume we know the height of persons (in chapter 4 we describe additional knowledge, e.g. door height, which can help to estimate a person’s height). The height of the bounding box in pixels is proportional to people’s height and distance from the camera. In general, we know that a real extent of a pixel is growing with distance.

Connecting the upper left corners of bounding boxes (or any corresponding ones) evaluated with the relative extent of pixels (from the known height of persons we can deduce the distance) can give us an estimation of the mutual distance between persons. This heuristic is very naïve and can give us only relative information on a person's mutual distance.

The second optical method is more involved and can serve as a baseline for the evaluation of experiments. The idea is that ground tags create a rectangular coordinate system. The perspective of the image creates a point where all perspectives intersect (assuming all is linear, recall all is only an illustration). The lower right corner of bounding boxes defines a rectangle ABCD (in green). We assume the distance of AD (=BC) can be estimated from the ground tags (yellow parallel lines, again a simplification assuming our camera preserves parallel lines perpendicular to the direction of the hallway). The distance DC (=AB) can be calculated from the known distance of a second row of ground tags (hidden behind person 2) and assuming that distances can be calculated linearly along lines perpendicular to the direction of ground tags (direction of walls of the hallway). Here we do not go into all peculiarities of imaging and camera properties.

The main idea is that this can be algorithmized. The method has to first recognize lines where the floor meets the wall (and we get the perspective). This has to be done only once as the camera is fixed. Second, the method has to recognize ground tags (maybe using some coding) and create a parallel line perpendicular to the direction of the hallway. The end position of the bounding boxes has to be aligned with this coordinate system.

Arbitrary images can be checked by human-mimicking construction. Moreover, having a scenario in which the real distance between persons in time when they stay on the ground tag is known can be also compared to the results of this method.

IV. Environment Knowledge

We discuss here additional environment knowledge which can enhance the transferability of our tool. We are restricted to indoor usage of our tools. This means that a potential customer can have a detailed plan of the environment with distances (possible distribution of furniture can be helpful).

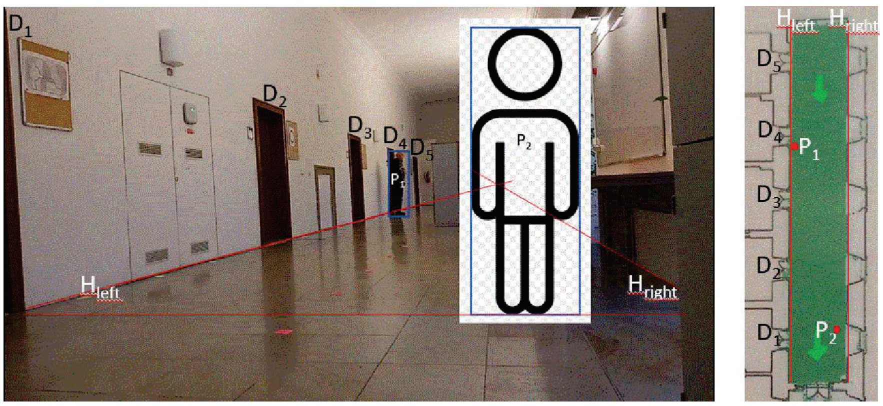

Fig. 2 An illustrative image of the identification of a camera shot and the hallway plan

As already mentioned, the height of doors can help to estimate the height of persons. In Figure 2 we see a possible illustrative plan of the hallway we depict. One can use the known positions of the door to estimate the position of the person in the hallway plan. Then the distance between persons can be calculated from a known scale of the plan. We also identify H_{left} and H_{right} with corresponding lines on the plan of the hallway.

It is a challenge for future work to design models and methods to create a correspondence between image and plan. For a fixed camera the task can be easier. The hope is that this knowledge can make adoption to a new environment easier. For each new customer, the model can be different, shape need not be rectangular. We hope that some basic information will be the same.

It is also a challenge for future work on how this correspondence between the image and the environment plan will be represented and used in the installation of our tool in a new environment. That is, if we assume we have image data annotated by object distances it should from the very beginning use some representation of the environment.

This is important in the case when only video data are used.

V. LOCALIZATION SENSORS

Having localization sensors changes the whole situation dramatically. Here we give a short overview of the use of UWB for the creation of video data annotated with object distance.

For distance estimation, we used the signal strength of Wi-Fi and Bluetooth first. Signal strength is proportional to distance and triangulation can give position and distances. We used NODE MCU ESP32 Wi-Fi + Bluetooth[3] as a scanner and MikroTik TG-BT5-IN[4] as tags. The first results were not conclusive. So, we decided to try also another direction. This does not mean we will not use BLE technology at all. We are just going in two directions in parallel and we will see to which ends it can bring us – especially with a synthesis of both.

Now we are experimenting with five UWB sensors (we used ESP32 board with integrated DWM1000[5], acting as both anchors and tags). The general idea of using UWB for positioning is in Figure 3, see[6].

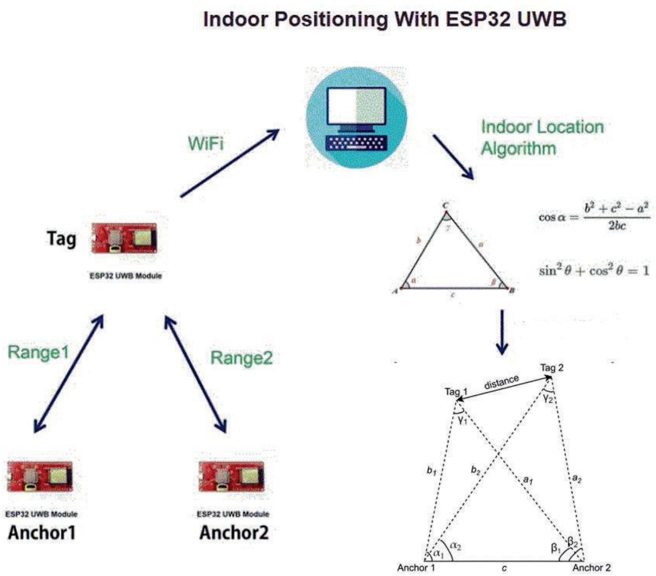

Fig. 3 An illustrative image of an indoor positioning system used for tags distance calculations (with https://www.instructables.com/ESP32-UWB-Indoor-Positioning-Test/)

Five sensors make it possible to have either 2 anchors and 3 tags (in a 2D projection of distance) or 3 anchors and 2 tags (for 3D in larger, mainly industrial, premises).

Data obtained can have the following form as in Table 1 (here we assume that distances of a tag to anchor 1 and anchor 2 come at the same time).

Table 1
An illustrative structure of data flow from two tags, two anchors and a camera

$TimeTag_1$	$dist_{a1}^{t1}$	$dist_{a2}^{t1}$	$TimeTag_2$	$dist_{a1}^{t2}$	$dist_{a2}^{t2}$	TimeCam	objectdist
$Time_1$	D_{time11}	D_{time12}					
			$Time_2$	D_{time21}	D_{time22}		
						$Time_3$	D_{time3}
$Time_4$	D_{time41}	D_{time42}					
			$Time_5$	D_{time51}	D_{time52}		
						$Time_6$	D_{time6}
$Time_7$	D_{tim711}	D_{time72}					
			$Time_8$	D_{time81}	D_{time82}		
						$Time_9$	D_{time9}

For the tag1(person1) distance in time $Time_1$ to anchor 1 is D_{time11} and to anchor 2 is D_{time12}. At another time $Time_2$ data from tag 2 arrive (person 2) again giving distances D_{time21} and D_{time22} from respective anchors. This is not exactly what is depicted in Figure 3, because we have distances from tags to anchors at different times. Nevertheless, we can assume that objects (persons) are

[3] https://www.espressif.com/en/products/socs/esp32
[4] https://mikrotik.com/product/tg_bt5_in
[5] https://www.makerfabs.com/esp32-uwb-ultra-wideband.html
[6] https://www.instructables.com/ESP32-UWB-Indoor-Positioning-Test/

moving slowly (assume persons are walking) and so the change of their distance in time $Time_2$ - $Time_1$ is negligible. Moreover, data from sensors and the camera are intertwined. The time is synchronized, which means that data flow was initiated in the same server time zero, nevertheless, responses can arrive at different times. Moreover, mixing may not be regular. It is the subject of further experiments in future work.

There is probably a big difference if we require the system to run online or offline on recorded data. So far, our consideration describes only an offline system. So, in $Time_3$ we probably do not know the distance between recognized persons D_{time3} (just an image/frame was saved). This is because the outputs of the neural network need some time to delineate the boxes in this image and then the distance needs to be calculated.

VI. Discussion, conclusions, and future work

Just walking around the nature of application domains, having The True Light at hand we can observe many interesting phenomena. Out there is the real life of humans and we can try to fine-tune their flexible nature. In this sense, we proposed a system and hope that it has algorithmic nature and can be implemented (whether offline or online with some additional effort of solving the problem of response time to be acceptable). Special attention was paid to the verifiability (resp. falsifiability) of our distance annotated data. Maybe the proposed controlled experiment is not sufficient and we have to add some additional features (e.g. recording sound, besides ground floor tags maybe some vertical tags on a wall can improve the controllability of the experiment). We experimented also with more cameras but the problem was in synchronization.

Although the COVID-19 pandemic is over we want to be prepared for future challenges in cooperation of healthcare and information technology.

Acknowledgment

This publication was realized with support of the Operational Program Integrated Infrastructure in frame of the project: Intelligent technologies for protection of health-care personnel in the front line and operation of medical facilities during spreading of disease Covid-19, code ITMS2014+: 313011ATQ5 and co-financed by the Europe Regional Development Found.

References

[1] Brezani S., Hrasko R., Vojtas P., Smart extensions to regular cameras in the industrial environment, Procedia Computer Science 200 (2022) 298-307, Available online 8 March 2022, In 3rd International Conference on Industry 4.0 and Smart Manufacturing, Edited by F. Longo, M. Affenzeller, A. Padovano, Elsevier, 2022, presented at ISM'21, 17-19 November 2021, Linz, https://www.msc-les.org/ism2021/.

[2] Brown, T. Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation. Harper Business (2009).

[3] J. Brownlee. How to Update Neural Network Models With More Data. Deep Learning March 5, 2021, https://machinelearningmastery.com/update-neural-network-models-with-more-data/ B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.

[4] T-Y. Lin et al. (2015) Microsoft COCO: Common Objects in Context. arXiv.org > cs > arXiv:1405.0312

[5] A. Kaklauskas et al Eds.. Lessons from Covid-19. Academic Press, 2022.

[6] P. H. Kindt et al. How Reliable Is Smartphone-Based Electronic Contact Tracing for COVID-19? Communications of the ACM, January 2022, Vol. 65 No. 1, Pages 56-67.

[7] https://www.ibm.com/ibm/history/ibm100/us/en/icons/servicescience/ The Invention of Service Science

[8] Dominic Tarn. How to Outsource Data Labeling for Machine Learning. Published November 11, 2022, Edited March 17, 2023, https://encord.com/blog/guide-to-outsourcing-data-labeling-for-machine-learning/

[9] Zhi-Hua Zhou and Ming Li, "Tri-training: exploiting unlabeled data using three classifiers," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529-1541, Nov. 2005, doi: 10.1109/TKDE.2005.186.