engineering proceedings

Proceeding Paper

Alternatives for Locating People Using Cameras and Embedded AI Accelerators: A Practical Approach [†]

Ángel Carro-Lagoa *, Valentín Barral, Miguel González-López, Carlos J. Escudero and Luis Castedo

CITIC Research Center & Department of Computer Engineering, University of A Coruña, 15071 A Coruña, Spain; valentin.barral@udc.es (V.B.); miguel.gonzalez@udc.es (M.G.-L.); escudero@udc.es (C.J.E.); luis.castedo@udc.es (L.C.)

* Correspondence: angel.carro@udc.es

[†] Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: Indoor positioning systems usually rely on RF-based devices that should be carried by the targets, which is non-viable in certain use cases. Recent advances in AI have increased the reliability of person detection in images, thus, enabling the use of surveillance cameras to perform person localization and tracking. This paper evaluates the performance of indoor person location using cameras and edge devices with AI accelerators. We describe the video processing performed in each edge device, including the selected AI models and the post-processing of their outputs to obtain the positions of the detected persons and allow their tracking. The person location is based on pose estimation models as they provide better results than do object detection networks in occlusion situations. Experimental results are obtained with public datasets to show the feasibility of the solution.

Keywords: indoor localization; computer vision; neural networks; embedded devices

check for updates

Citation: Carro-Lagoa, Á.; Barral, V.; González-López, M.; Escudero, C.J.; Castedo, L. Alternatives for Locating People Using Cameras and Embedded AI Accelerators: A Practical Approach. *Eng. Proc.* **2021**, *7*, 53. https://doi.org/10.3390/engproc2021007053

Academic Editors: Joaquim de Moura, Marco A. González, Javier Pereira and Manuel G. Penedo

Published: 27 October 2021

1. Introduction

During recent years, the interest in indoor human location has increased due to the large number of applications in various fields, such as security surveillance, activity monitoring, behavioral analysis, and healthcare [1].

Traditionally, when it is necessary to locate people indoors, radio-based technologies are used, which can be affected by the characteristics of the environment and also force target users to carry specific devices. An alternative to these technologies is video-based localization using security cameras, which are increasingly common in buildings and public places. Due to advances in computer vision and Deep Learning (DL), the detection of people on video is more reliable.

The localization and tracking of people is usually performed in two steps: people are first detected in each individual frame to obtain their position in the image. Then, these detections are associated across frames to obtain the path followed by each person.

Typically, these tasks are performed by processing the video from each camera in a centralized way. However, it is possible to perform this processing in a distributed manner due to the advances in edge-computing. Cameras can use AI accelerator chips that allow for fast and low-power neural network inference. Several chips are available on the market, such as Google Coral, Intel Movidius, or Nvidia Jetson.

In this work, we focus on the task of detecting people in security camera images by performing the processing on an embedded device with a Google Coral's Edge TPU. The tracking methods that can be applied to the obtained results are not addressed.

2. Person Location Method

Pedestrian location is performed by processing the input images with a person detector that will locate the persons present in the image. Then, the output of the detector is processed to determine the real-world position of each person. To perform this last step, the camera calibration information and the vertical vanishing point are also used.

2.1. Person Detection

The most common alternatives for person detection using convolutional neural networks (CNNs) are object detectors, which provide bounding boxes of the persons, and pose estimation, which provides the position of the different key body joints of each person. Cosma et al. [2] compared these two methods, obtaining better results with pose detection networks, which are more resistant to occlusions. Moreover, the correct processing of the detected pose allows for estimation of the position of the person's feet more accurately even when they do not appear in the image.

The PoseNet neural network [3] is used in this work.This model uses a bottom-up approach where all the keypoints of every person are first predicted using a CNN, providing a heatmap for every body part. Then, these keypoints are grouped into individuals using a custom greedy algorithm. This last step can fail if the image has several persons close to each other, mixing the keypoints of two or more persons.

There are several pretrained PoseNet networks available with different CNN backbones and input resolutions. We selected the ResNet50 backbone with a 416 × 288 resolution as it provides a good balance between inference speed and reliability.

2.2. Post-Processing of Person Keypoints and Projection to 2D Map

The keypoints of each person are used to predict the position of the feet, even if they are not detected or they are occluded. Each keypoint obtained has a score, allowing discarding the keypoints with low reliability. With these reliable keypoints, our post-processing algorithm predicts the feet and head position of each person taking into account the proportions of the human being. These positions are estimated using the least squares method. The vertical vanishing point of the image is also taken into account to correct the inclination of people in the image, depending on the camera perspective.

Cosma et al. [2] used a similar method with the following differences: they only performed these calculations when the feet positions were not detected, and they attempted to determine the inclination of people without taking into account the vanishing point.

Once the feet position in the image is known, the information from the camera is used to determine the map position of each person. The correspondence between each pixel on the image and the 2D floor map coordinates can be calculated with a homography transformation. The homography matrix can be obtained from the position of, at least, four pixels and the map coordinates of each of them. This matrix can also be calculated from the camera projection matrix.

In certain situations, when a person is very close to the camera and only the head is detected, the estimation of the feet position is very poor. This problem can be corrected by assuming that the person has an average height and using the known position of the camera, thus, providing a better estimation of the person's position.

3. Experimental Results

The CamLoc [2] and ICG Lab6 [4] datasets were processed with our person positioning system. Unlike other datasets that only provide the bounding boxes of each person, these datasets annotate the groundtruth position of each person in the map and provide the camera calibration information. This enables us to directly obtain the mean error of the estimated positions.

The CamLoc dataset contains only one person in several scenarios with varying levels of occlusion. Table 1 shows the obtained results with the CamLoc dataset. The mean error of the positions and the percentage of missing predictions are compared, showing that our system obtained better results with all the cameras.

Table 1. The results with the CamLoc dataset compared with the original results.

Camera	Mean Error (cm)		Missing Predictions (%)	
	CamLoc	This Work	CamLoc	This Work
S1_Wide_cam1	36.26	27.49	9.18%	3.46%
S1_Wide_cam4	53.58	45.39	4.47%	0%
S2_Narrow_cam2	45.27	30.96	-	5.72%

The ICG Lab6 dataset [4] consists of one room that is simultaneously recorded by four cameras. There are six scenarios where several persons perform different activities in the room. Table 2 shows the obtained results jointly with the results in [4]. In addition to the mean error, the detected true positives (TP), false positives (FP), and false negatives (FN, i.e., the missing detections) are shown.

Table 2. The mean error column only considers the error of the TP detections.

Scenario	Algorithm	Mean Error (m)	TP	FP	FN
CHAP	ICG Lab6	0.102	1555	2	6
CHAP	This work	0.105	1513	33	25
LEAF1	ICG Lab6	0.107	464	2	2
LEAF1	This work	0.092	422	8	39
LEAF2	ICG Lab6	0.097	930	41	41
LEAF2	This work	0.102	517	15	453
MUCH	ICG Lab6	0.111	783	9	9
MUCH	This work	0.098	780	28	10
POSE	ICG Lab6	0.123	485	14	14
POSE	This work	0.150	428	31	57

The ICG Lab6 method uses a specific tracking algorithm for this kind of scenario with several cameras covering a common area and obtains good results. Our results were obtained by performing a merge of the near positions detected in each camera, and then using a simple tracking algorithm to filter out some FP and FN, only considering the positions of the detections and not the appearance of each person. Moreover, the results are also affected by the difficulty of the pose estimator to distinguish between people when they are very close to each other.

4. Conclusions

We described the developed person location method based on computer vision techniques and provided our experimental results. The obtained results showed the high accuracy that this kind of positioning system can provide. However, in complex scenarios, an adequate tracking algorithm that takes into account the appearance of each person is needed to obtain reliable results.

Funding: This work has been funded by the Navantia-UDC Joint Research Unit under Grant IN853B-2018/02, the Xunta de Galicia (by grant ED431C 2020/15, and grant ED431G 2019/01 to support the Centro de Investigación de Galicia "CITIC"), the Agencia Estatal de Investigación of Spain (by grants RED2018-102668-T and PID2019-104958RB-C42) and ERDF funds of the EU (FEDER Galicia 2014–2020 & AEI/FEDER Programs, UE).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Morar, A.; Moldoveanu, A.; Mocanu, I.; Moldoveanu, F.; Radoi, I.E.; Asavei, V.; Gradinaru, A.; Butean, A. A Comprehensive Survey of Indoor Localization Methods Based on Computer Vision. *Sensors* **2020**, *20*, 2641. [CrossRef] [PubMed]
2. Cosma, A.; Radoi, I.E.; Radu, V. CamLoc: Pedestrian Location Estimation through Body Pose Estimation on Smart Cameras. In Proceedings of the 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Pisa, Italy, 30 September–3 October 2019; pp. 1–8. [CrossRef]
3. Papandreou, G.; Zhu, T.; Chen, L.C.; Gidaris, S.; Tompson, J.; Murphy, K. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
4. Possegger, H.; Sternig, S.; Mauthner, T.; Roth, P.M.; Bischof, H. Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.