

People detection with omnidirectional cameras using a spatial grid of deep learning foveatic classifiers

Daniel Fuertes^{a,*}, Carlos R. del-Blanco^a, Pablo Carballeira^b, Fernando Jaureguizar^a, Narciso García^a

^a Grupo de Tratamiento de Imágenes (GTI), Information Processing and Telecommunications Center, ETSI Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain

^b Video Processing and Understanding Lab, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain

ARTICLE INFO

Article history:

Available online 18 February 2022

Keywords:

Spatial grid
Deep learning
Omnidirectional cameras
People detection
Point based annotations

ABSTRACT

A novel deep-learning people detection algorithm using omnidirectional cameras is presented, which only requires point-based annotations, unlike most of the prominent works that require bounding box annotations. Thus, the effort of manually annotating the needed training databases is significantly reduced, allowing a faster system deployment. The algorithm is based on a novel deep neural network architecture that implements the concept of Grid of Spatial-Aware Classifiers, but allowing end-to-end training that improves the performance of the whole system. The designed algorithm satisfactorily handles the severe geometric distortions of the omnidirectional images, which typically degrades the performance of state-of-the-art detectors, without requiring any camera calibration. The algorithm has been evaluated in well-known omnidirectional image databases (PIROPO, BOMNI, and MW-18Mar) and compared with several works of the state of the art.

© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

People detection is a very active field due to an endless number of industrial and commercial applications: analysis of pedestrian flows, defense and security, interaction for intelligent buildings, immersiveness for virtual/augmented environments, geofence, etc. There are different technologies for people detection in indoor environments: radio-frequency identification (RFID), wireless local area networks (WLANs) [1,2], ultra wide band (UWB) [3], Bluetooth [4,5], inertial measurement unit (IMU) [6], magnetic field [7], or even cloud-based approaches that combine several technologies [8]. All these technologies require the use of active sensors, in such a way that a person can only be located if he/she carries such a sensor (or combinations of them embedded in a mobile phone), and agrees to share his/her data. Alternatively, systems based on passive cameras have the potential to detect and locate people without the previous restrictions.

Most of the current people detection systems use standard perspective projection cameras with a limited field of view that can only cover reduced areas. A solution to monitor large areas is to

install a network of cameras, but it significantly increases the deployment cost of the system. Omnidirectional cameras can be also used to cover large areas, taking advantage of their very wide field of view that can reach over 180 degrees. Thus, one omnidirectional camera is enough for environment monitoring, reducing the system cost. However, omnidirectional imagery undergoes strong geometric distortions, which cannot be handled by most of the existing people detection algorithms, usually developed for perspective projection cameras.

On the other hand, the notable improvement of the detection capabilities of the current people detection systems is greatly due to the adoption of machine learning techniques, which make use of prior knowledge in form of annotated databases to boost the detection performance [9,10]. Of special interest are the deep learning techniques, a subfamily of machine learning ones, which are end-to-end trainable, i.e., fully optimizable, and have reported important performance improvements in a wide range of applications. These techniques involve a training stage that requires huge datasets to make accurate detections. The creation of such datasets is often a remarkable problem in the deployment of a practical system, due to the high time and monetary cost in their generation. The most problematic part is the manual generation of the upright bounding-box annotations per frame. Typically, this process requires that a human manually annotates two points (which defines the bounding box) per person and image, continuously

* Corresponding author.

E-mail addresses: d.fcoiras@upm.es (D. Fuertes), carlosrob.delblanco@upm.es (C.R. del-Blanco), pablo.carballeira@uam.es (P. Carballeira), fernando.jaureguizar@upm.es (F. Jaureguizar), narciso.garcia@upm.es (N. García).

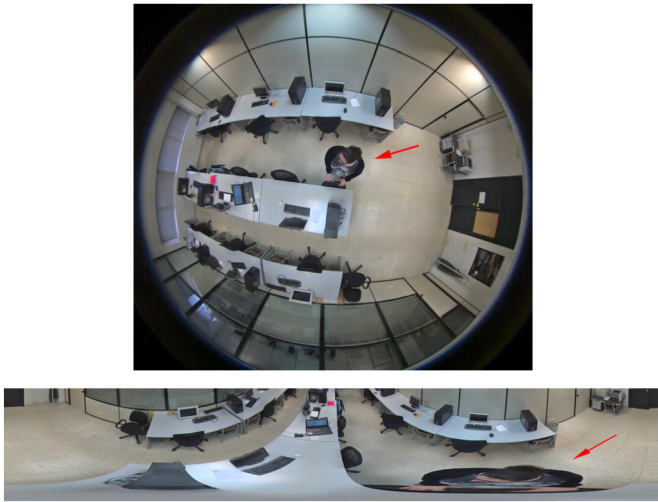


Fig. 1. Upper image: omnidirectional image showing the geometric distortions induced by a wide-angle lens. Bottom image: unwarping of the previous omnidirectional image to a perspective-like one, where several typical artifacts can be observed.

accommodating changes in size and position along a video sequence. The situation is even worse for omnidirectional imagery since the inherent geometric deformations make the traditional upright bounding boxes not very suitable for that purpose. And, the typical approach of unwarping an omnidirectional image to obtain a perspective-like one is far to be perfect, as shown in Fig. 1. For this challenge, some works [11] have proposed to annotate rotated bounding boxes so that they can encompass better the people inside. However, it implies a higher annotation effort. Additionally, the bounding boxes (either upright or rotated) undergo larger variations in position and size along time, increasing the amount of annotation work.

The previous annotation problem would be largely alleviated if a machine learning system could accept just points as annotations, each representing a symbolic reference of a person (such as the chest or the head). Thus, the annotation effort is just reduced to one point per person, instead of two. But also, the annotation would be easier and faster since there would be no need of considering the temporal changes in shape and size.

For this purpose, a new human detection system based on omnidirectional cameras and deep learning techniques is proposed in this paper, called Grid of Spatial-Aware Classifiers based on Deep Neural Networks (GSAC-DNN), which can be trained with just point-based annotations. The system can directly deal with the distortion of the omnidirectional imaging without requiring any unwrapping stage or camera calibration. The proposed deep neural network architecture is based on the concept of the Grid of Spatial-Aware Classifiers (GSAC), presented in [12,13], but, unlike the previous work, GSAC-DNN is end-to-end trainable, allowing a jointly optimization of the whole system. In more detail, the GSAC-DNN design is implemented as a neural network architecture by a 2D array of neural network branches, which are composed by a convolutional and a fully connected layer. These branches share the same feature tensor as input, computed by a Convolutional Neural Network (CNN) backbone that processes every image. This multi-branch-classifier structure adapts very well to the distortion of omnidirectional images, since every classifier branch is in charge of learning a subset of the severe appearance changes of people induced by the omnidirectional geometry distortion, which in turn depends on the specific image position. Last but not less, the proposed algorithm can work in a graphics processing unit (GPU) in real-time.

The organization of the paper is as follows. The state of the art in people detection using omnidirectional cameras is reviewed in Section 2. Section 3 presents the people detection proposal based on a deep neural network architecture. Section 4 evaluates and compares the proposed algorithm in several datasets and state-of-the-art works. Finally, the conclusions are provided in Section 5.

2. State of the art

People detection methods working on omnidirectional images can be split into two different subgroups: methods that perform some image adaptation and methods that directly work on the omnidirectional domain. Methods performing image adaptations, also referred as to unwarping methods, include a pre-processing stage in which omnidirectional images are geometrically transformed into perspective-like ones, as shown in Fig. 1. Thus, detection algorithms originally conceived for conventional cameras can be applied to omnidirectional ones. Unwarping methods require to estimate the calibration parameters of the camera, making the detection system deployment more complex. On the other hand, the unwarping can be performed at image region level [14,15] or to the whole omnidirectional image. Sheng *et al.* [16] proposed an unwarping stage at image region level that focuses on reducing the computational cost caused by the redundant overlapping among the image regions to be unwrapped. They perform the detection task on several concatenated perspective views, instead of unwarping individual views from the fish-eye camera, using a neural network with a YOLOv3 [17] architecture, a specialized CNN for object detection. The unwarping can be alternatively performed to the entire omnidirectional image, either to achieve a perspective-like image or a panoramic one. In [18], perspective images are obtained from the whole omnidirectional images, which are then processed by the previous version of YOLOv3, that is, by YOLOv2 [19]. In [20,21], fish-eye images are geometrically transformed into 360-degree panoramic images for both human detection and human action recognition. However, the unwarping of omnidirectional images is not perfectly possible in practice, containing errors or artifacts that inevitably affect the detection performance [12].

Methods that directly work on the omnidirectional domain can be divided into two subfamilies: feature adaptations methods and those ones learning directly on the omnidirectional domain. The first ones adapt existing feature extraction techniques, originally designed to deal with perspective images, to work on omnidirectional ones. In [22], Aggregate Channel Features (ACF), originally designed for perspective images, are directly computed from omnidirectional images, and then post-processed to obtain perspective-like features. This strategy allowed to just use a classifier trained on perspective images for the purpose of people detection. In [23], Histograms of Oriented Gradients (HOG) [24] features are computed from rotating sliding windows (adopting a circular sector shape) for the human detection task. Although an improvement was achieved regarding the image unwarping strategy, the designed feature adaption was only capable of managing a restricted set of human appearance distortions of the omnidirectional images. In [25], a procedure to create a database simulating some of the geometric distortions of the omnidirectional images is proposed. The database is then used to train two different CNNs based on YOLOv2 and YOLOv3 architectures with the purpose of detecting humans in omnidirectional images. Although not specifically related to people detection or to omnidirectional cameras, a rotation-invariant CNN is proposed in [26], which could be applied to manage some of the omnidirectional image distortions. Similarly, a combination of regularizers for promoting rotation-invariance is used for training a CNN in [27]. The other family of methods, which learn directly from omnidirectional images, train a classifier using annotations over the own omnidirectional images.

In [28], the Mask-RCNN algorithm is fine-tuned with a reduced set of fish-eye images via a transfer-learning approach. However, a more exhaustive training using large datasets is still desirable to improve the performance of this kind of large neural-network architectures. This is somewhat alleviated in [29,30] by proposing less complex neural networks with fewer parameters. Specifically, Tiny YOLOv1 and another lightweight version of YOLOv1, called Simplify YOLOv1, are trained on omnidirectional images. They are used in combination of a foreground segmentation mask to try to improve the detection accuracy of moving people. However, it cannot locate stationary people by design. [31] faces a similar problem proposing two different methods: an Activity Blind (AB) method, based on rotating sliding windows around the center of the image that are then fed into YOLOv3 to perform detections, and an Activity Aware (AA) method that performs similar to AB but extracting regions of interest with background subtraction to reduce the number of sliding windows. However, AA is not capable of detecting stationary people, just like the previously mentioned Tiny YOLOv1 and Simplify YOLOv1, and AB has a very high computational cost because it is based on a sliding-window strategy. In [11], a deep learning system based on YOLOv3 modified to use rotated bounding boxes is proposed. It achieved a better accuracy thanks to the fact that the rotated bounding boxes could encompass better the person silhouette. However, a higher cost was required for the annotation task. On the other hand, this kind of works based on convolutional neural networks inherently assumes that the visual appearance of the objects is approximately invariant to translations. However, this is not true in omnidirectional images, where object appearance undergoes large changes according to its spatial location in the image. As a result, the recognition capability of the system could be degraded to some extent.

As a common fact, all the previous approaches, especially those based on deep neural networks, require huge annotated databases to properly make predictions. The annotations are usually based on (upright) bounding boxes, whose manual generation implies a high cost in time and money. Even the situation is worse for those systems that require rotated bounding boxes since they require more effort. Some proposals have been made to reduce the annotation burden. For example, in [32], a weakly supervised learning approach based on a Bayesian framework is used to detect objects in optical remote sensing images. Another strategy that minimizes the annotation effort is proposed in [33] for the task of hand gesture recognition, which only uses one label per image sequence, instead of one bounding box per frame to enclose every hand instance in the sequence. However, they either do not adapt to the specific problem of people detection or cannot be adapted to an end-to-end trainable neural network framework.

Interestingly, the previous problems could be alleviated, and even solved, by using the main ideas of the detector presented in [12] for people and later adapted for vehicles in nighttime in [13]. This detector proposes to use a pool of classifiers covering different spatial regions in the image (not necessarily exclusive), where the precise monitored area by each classifier is automatically learned according to the expected visual appearance of the objects in each image location. Thus, each classifier is specialized in recognizing the people's appearance in a certain spatial image location, which are largely influenced by the geometrical distortions of the omnidirectional camera lenses. As a result, the GSAC detector can adapt to the special characteristics of omnidirectional images. Additionally, it can be trained with just point-based annotations, which not only reduces the annotation effort to one point per person (instead of two or three), but also it avoids tracking changes in size and shape, making the annotation process much easier and faster. However, the original implementation included several independent modules that cannot be globally optimized, i.e., it is not an end-to-end trainable system, and therefore, it can-

not benefit from the proven boost of performance of those works that are fully trainable. More specifically, a pool of independent Support Vector Machines (SVM) was used in GSAC to process a common image feature vector based on a variation of the HOG descriptor.

In this paper, a novel deep neural-network architecture that implements the ideas behind the GSAC detector is proposed for the task of real-time people detection in omnidirectional images. The proposed system, GSAC-DNN, includes several contributions with respect to GSAC. First, it can be fully optimized via an end-to-end training procedure, while GSAC required to train independently each SVM classifier. Second, it employs the backbone of the network to automatically extract feature maps adapted to maximize the application performance, instead of using handcrafted HOG features. Each classifier from the grid is equipped with a convolutional layer that allows to personalize the feature maps for the specific conditions of the classifier. Moreover, the system can be executed on a central processing unit (CPU) or a GPU without a specific implementation for every hardware. Regarding the previous approaches that require bounding box annotations (enclosing the person to be detected), the proposed neural network architecture uses the same strategy for point-based annotations (representing a characteristic human body part) as GSAC. Thus, the database annotation cost is significantly reduced, and the system can be deployed faster.

3. Description of the detection system

The proposed detection system for omnidirectional imagery, called GSAC-DNN, is based on a novel deep neural network with multiple output branches that implements the design ideas of the GSAC detector. It presents three remarkable characteristics: 1) it is fully optimizable via an end-to-end training procedure; 2) it can be trained with just point-based annotations; and 3) it can cope with the geometric distortions of omnidirectional imagery that causes strong variations in the person's appearance. The block diagram of the GSAC-DNN detector is depicted in Fig. 2. A feature map per image, f_I , is computed using ResNet-32 [34], a well-known CNN backbone that offers a high convergence capability for the training stage via residual connections. The resulting feature map f_I is processed in parallel by a 2D grid of neural network branches, $C = \{C_i | i = 1, \dots, N_C\}$, every one representing a spatial-aware classifier implemented as the concatenation of a convolutional layer and a fully connected layer. The output of the whole set of branches is a 2D grid of independent binary outputs (one per branch). The common branch architecture is shown in Fig. 3, which is composed of two layer blocks. The first one contains a convolutional layer, a batch normalization layer, and a ReLU function. The second block contains a fully connected layer followed by a sigmoid activation function that implements a binary output (the presence or not of a person). Each spatial-aware classifier, C_i , has associated a 2D spatial point over the image, which is used as reference to detect only the people in its surroundings. Note that the set of spatial reference points related to the classifiers can be represented by a 2D grid over the omnidirectional image (see Fig. 4). Thus, every classifier branch deals only with a subrange of the changes in the people's appearance, largely caused by the geometric deformations of the omnidirectional imagery. This performance is boosted by the convolutional layer of every classifier, which will produce specific feature maps adapted to the position of the classifier in the image. The output of C is a 2D grid of confidence scores reporting about the presence or not of a person in the neighborhood of every spatial reference related to a classifier (the called awareness area). This architecture allows to optimize together both the weights of the CNN feature extractor and those in the binary classifiers, i.e. it can be trained end-to-end.

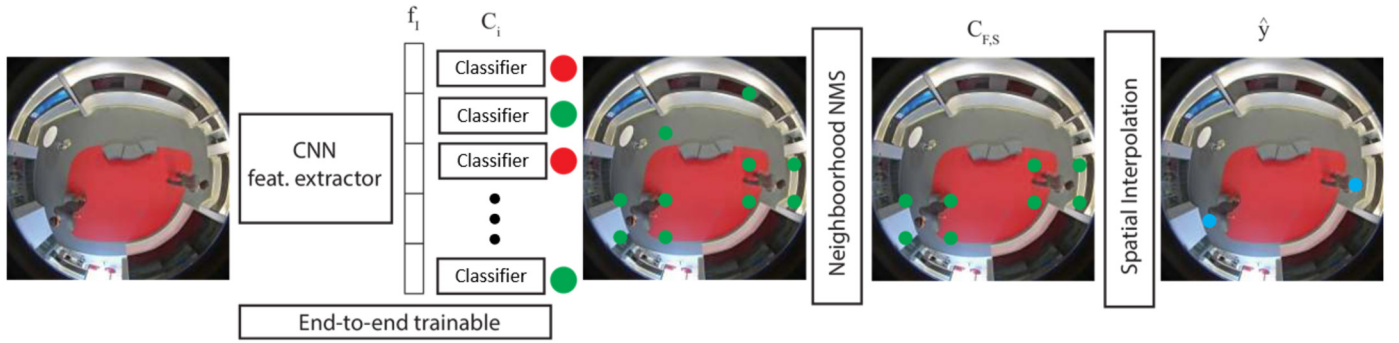


Fig. 2. Block diagram of the GSAC-DNN detection system. The circles represent the predictions of the grid of classifiers. Green circles correspond to active classifiers. Red circles correspond to inactive classifiers that, for clarity, have been omitted in the images. And blue circles represent the final estimated locations. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

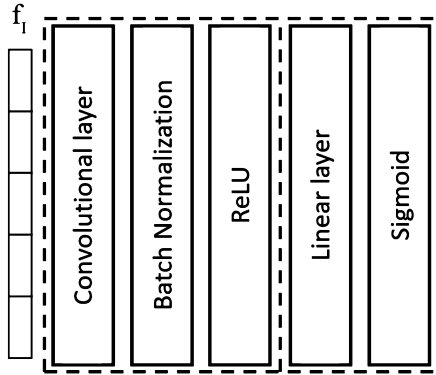


Fig. 3. Common architecture of the neural network branches representing the 2D grid of spatial-aware classifiers.

The final location of people is obtained after a non-maxima-suppression (NMS) and a spatial interpolation (to achieve sub-grid precision). This process is outlined in Fig. 2 and depicted in detail in Fig. 4. First, the outputs of C , containing the confidence scores of every classifier in the 2D grid, are refined by removing low score values (related to putative false detections). This filtering procedure is carried out by computing a neighborhood-based score per classifier C_i , where confidence scores are averaged inside a given neighborhood C_U . Then, classifiers with a low confidence neighborhood score are filtered out, obtaining C_F . Then, non-maxima suppression is applied to C_F to discard more than one detection per person, and thus decreasing the false alarms, obtaining $C_{F,S}$. Lastly, an accurate estimation of the location \hat{y} of each person is obtained by linear interpolation of the spatial locations of every neighborhood in $C_{F,S}$.

3.1. Online labeling for the training stage

The training of GSAC-DNN implies an online conversion of the initial point-based annotations into arrays of binary labels that represent the output of the grid of classifiers. The point annotations of any image, containing an arbitrary number of point-based annotations (or even no annotations), is converted to an array of binary labels of size N_C . The specific binary value of each array element is computed as a function of the distance between each ground-truth point annotation and the spatial references of all classifiers. Specifically, N_p classifiers are activated, those ones whose reference points are the closest to the location marked by the point-based annotation, while the others are deactivated (set to zero). Figs. 5 and 6 illustrate this procedure.

4. Evaluation and results

4.1. Datasets and metrics

Three public datasets containing omnidirectional imagery have been used to assess the detection accuracy of GSAC-DNN, and make a comparison with other relevant works in the state of the art: PIROPO [12], BOMNI [35] and MW-18Mar [36]. The PIROPO database is composed of multiple sequences acquired in two rooms (Fig. 7). The first room (A) is equipped with three omnidirectional cameras located on the ceiling, with an image resolution of 800×600 pixels. The other one has only one omnidirectional camera, also on the ceiling with the same characteristics. The annotations are point-based (each person is annotated by a point in his/her head), reaching the amount of more than 100,000 manually annotated frames. Additionally, there exists partial extended annotation of the database in form of bounding boxes provided by other works [16][15]. BOMNI is a smaller dataset composed of sequences acquired from two omnidirectional cameras located in the same room at the ceiling and on the wall. The original use of this database was the evaluation of people tracking algorithms, but it has also been used for people detection. But, only the sequences acquired by the camera that is mounted on the ceiling are used (Fig. 8), since the other one undergoes large occlusions most of the time because of being located on a vertical wall. The image resolution is 640×480 pixels and the annotations are given by bounding boxes. MW-18Mar is composed of seven omnidirectional ceiling cameras located on different indoors environments, as shown in Fig. 9. It has been used for different computer vision tasks, such as person identification, detection, and tracking. Depending on the sequence, the image size can be 1056×960 or 1488×1360 , both resolutions larger than the ones from PIROPO and BOMNI.

To test the proposed GSAC-DNN detection system, point-based annotations for the BOMNI and MW-18Mar databases are obtained from the bounding boxes, by computing the bounding box centroid.

Common object and people detection metrics have been used, which can be divided into two groups. The first one composed by Precision (P), Recall (R), F1-score (F), and Average Precision (AP) [37]; and the second one by Miss Rate (MR), False Positives Per Image (FPPI), and Logarithmic Average Miss Rate (LAMR) [38]. Before defining those metrics, it is necessary to categorize the predictions in: True Positives (TP) that informs about the number of persons that were correctly detected; False Positives (FP) that counts the number of wrong detections; and False Negatives (FN) that provides the number of missed people. Usually, the predicted coordinates do not exactly match those of the ground truth, but expectedly they are very close to them. To measure this misalign-

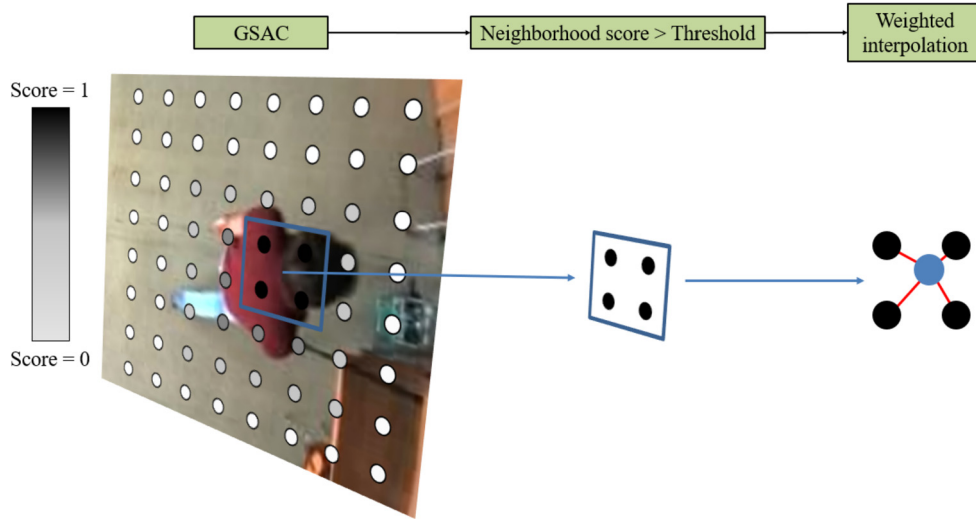


Fig. 4. Fusion of outputs of the classifier branches to obtain the final person location. A neighborhood of classifiers (circles) whose confidence score surpasses a determined threshold triggers the prediction of a person (blue circle marker), whose final location is obtained through a weighted interpolation of the output scores.

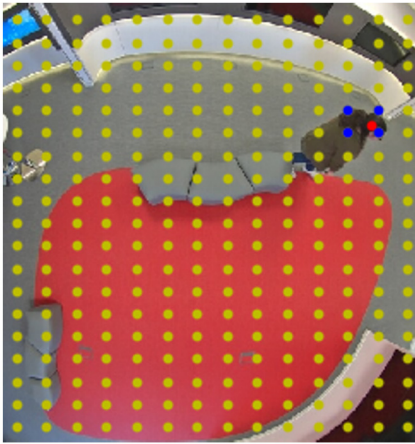


Fig. 5. Example of point-based annotation in red. In this case, the labels of the four classifiers that are closer (considering their reference points) to one point-based annotation are set to one, while the others are set to zero.

Positive and negative sample association

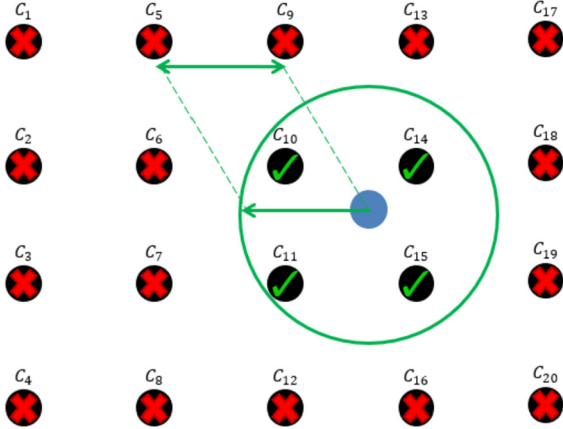


Fig. 6. Positive and negative sample association. The labels related to classifiers whose spatial references (black circles) are closer to each point based annotation (blue circle) are marked as one (green check mark), and the others as zero (red cross mark).

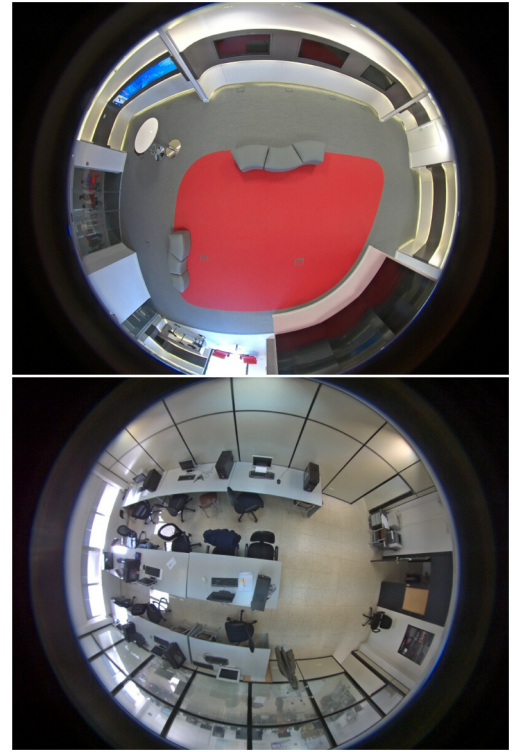


Fig. 7. Sample images: Rooms A and B of the PIROPO database.

ment in point-based annotations, a distance criterion based on the Euclidean distance has been used.

After estimating TP, FP, and FN, Precision and Recall can be defined. The Precision measures the effectiveness of the system at predicting correct people. If the precision is high, every time a person is predicted, it is probably correct. The Recall measures the ability of the system to detect all the people. A high recall value means that the system is very likely to notice all the persons. Precision and Recall can be expressed as:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad (1)$$



Fig. 8. Sample image of BOMNI database.

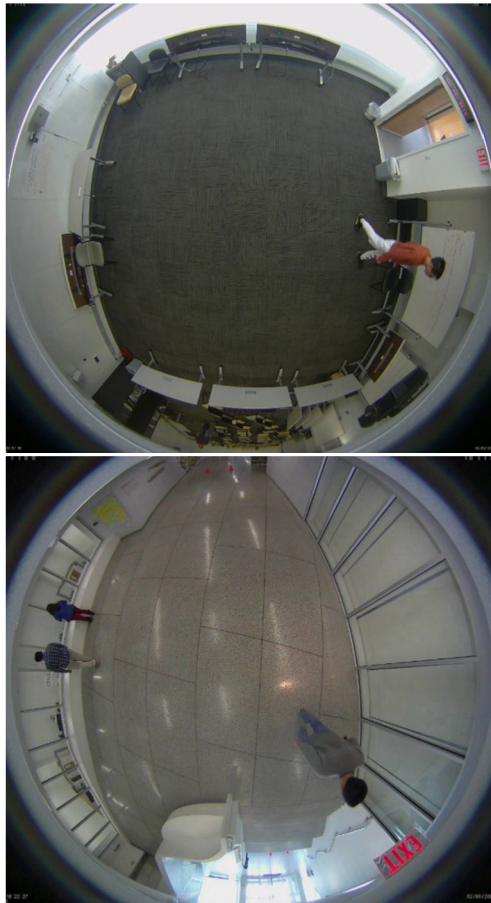


Fig. 9. Sample images of MW-18Mar database.

With Precision and Recall metrics, it is possible to define the Precision-Recall curve. It is computed by ordering the predictions according to their confidence score, and then calculating the Precision and the Recall for each. The area located below the Precision-Recall curve is called Average Precision, and it measures the balance between the Precision and the Recall under different working conditions of the detector.

Regarding the MR and FPPI, they can be considered as the inverse of the Precision and the Recall. They are defined as:

$$MR = \frac{FN}{TP + FP} \quad FPPI = \frac{FP}{N_{images}} \quad (2)$$

Table 1

Results for the PIROPO dataset using the metrics: Precision (P), Recall (R), F1-score (F), and Average Precision (AP).

PIROPO	P	R	F	AP
GSAC-DNN (ours)	0.988	0.925	0.956	0.919
GSAC [12]	0.928	0.878	0.901	0.887
Adapted YOLOv3	0.945	0.926	0.935	-
Adapted Faster-RCNN	0.960	0.900	0.929	-
Based on YOLOv3 [16]	-	-	-	0.798

where N_{images} is the number of test images. If the average of the MR is calculated at several FPPI rates evenly spaced in log-space, the LAMR metric is obtained, which is conceptually the inverse of the AP. Lastly, the computational complexity has been measured in Frames Per Second (FPS). Note that a global comparison for all the involved works is not possible, since every work only includes a subset of the previous metrics, and the code to reproduce the results is not available.

4.2. Evaluation results

A comparison of different people detection systems using the previous datasets and metrics is provided. Regarding GSAC-DNN, input images have been down-scaled to 224×224 pixels (due to limitations of the used GPU memory in the training procedure) and a grid of 28×28 classifiers has been used. Considering the size of the images and the grid of classifiers, the distance between classifiers is 8 pixels. Therefore, the threshold for the distance criterion based on Euclidean distance is fixed to twice the distance between classifiers, that is, 16 pixels, in order to ensure that the distance between all the neighbor classifiers is not larger than the threshold. For the training stage, Adam optimization with an initial learning rate of 10^{-3} has been adopted. The decay rate has been fixed to several values, 10^{-1} , 10^{-2} , 10^{-3} , and $0.5 \cdot 10^{-3}$, depending on the number of trained epochs. These parameter values have been obtained after an extensive experimentation and evaluation process (this implies evaluating multiple parameter configurations, analyzing their results and parameter dependencies, and then selecting what new range of parameters are the most promising to evaluate in the next trial. This process is repeated until no further improvement is obtained, selecting as the final parameters those that have reported the best performance for the system). Data augmentation has been also used by including random flippings and contrast and brightness changes. No pretrained weights have been used.

Some qualitative results are provided in Fig. 10. This Figure exposes less obvious situations that could propitiate a bad performance of the system. However, GSAC-DNN performs reasonably well for almost all the cases. Only at the bottom-right image, GSAC-DNN misses some detections because the people is very close to each other and two of them are partially occluded. The occluded people will offer very few visual information to the algorithm, while the fact that the people is close to each other implies that the NMS stage is very likely to consider some of the True Positives as False Positives.

The methods of the state of the art used for the comparison have been already commented in the Section 2. All of them used (upright) bounding boxes or rotated bounding boxes for the dataset annotations. To offer a more complete perspective, two additional baselines have been included in the comparison based on the detectors YOLOv3 [17] and Faster R-CNN [39], but adapted to use point-based annotations (like GSAC-DNN). The adaptation consists in taking several bounding boxes of different sizes centered in the original point-based annotations of PIROPO. Then, those bounding boxes are employed for the training of YOLOv3 and Faster R-CNN. From the multiple performed trainings (each one with a

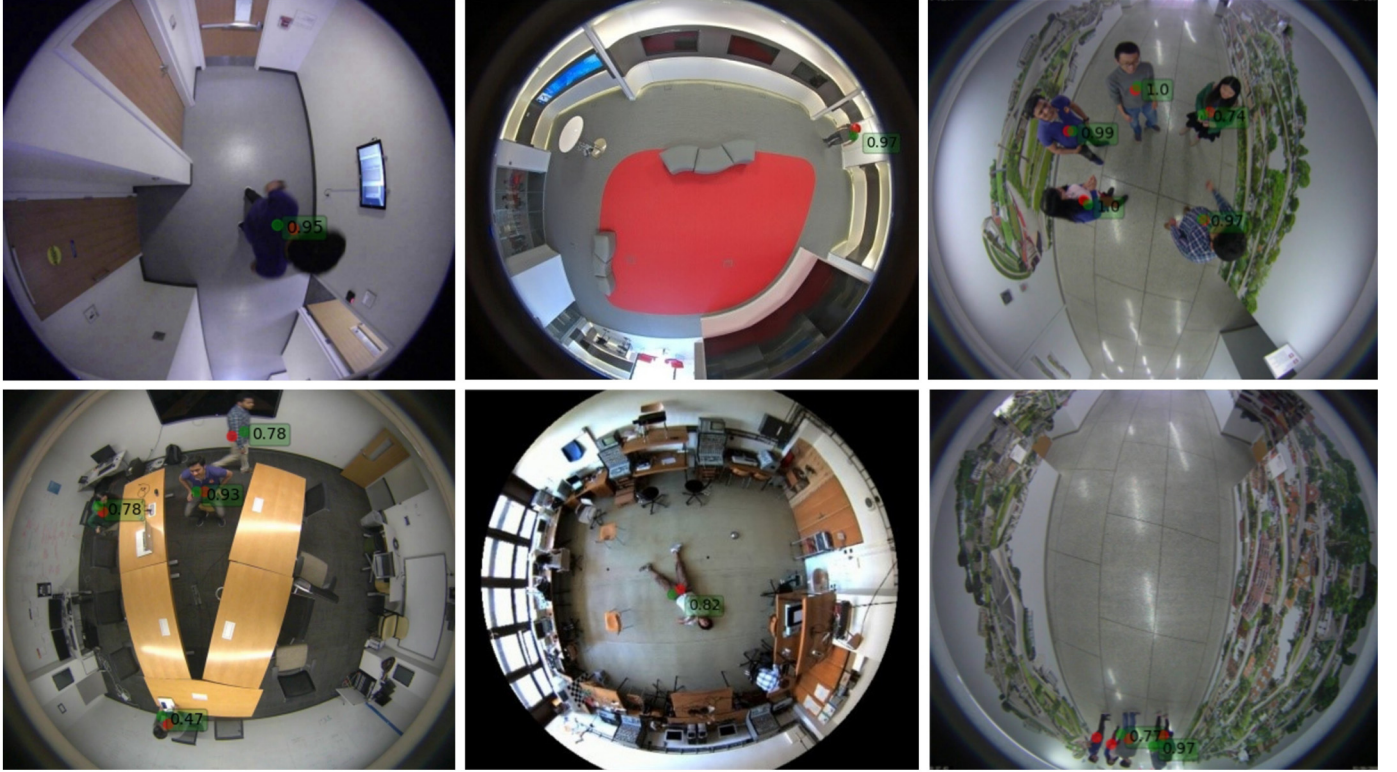


Fig. 10. Examples of less obvious situations on the three datasets (PIROPO, BOMNI and MW-18Mar), where a green dot refers to the predicted marker and a red dot, to the ground-truth. Upper images (from left to right) show examples of a person located at the center of the image, a person located at the periphery of the image and a small group of people. Bottom images (from left to right) show examples of persons sitting on chairs, a person laying on the floor and a group of persons where some of them occlude others.

Table 2

Results for the BOMNI dataset using the metrics: Precision (P), Recall (R), F1-score (F), and Average Precision (AP).

BOMNI	P	R	F	AP
GSAC-DNN (ours)	0.951	0.923	0.937	0.921
GSAC [12]	0.890	0.869	0.879	0.856
Unwrapped ACF [22]	0.800	0.850	0.824	-
Tiny YOLOv1 [30]	0.561	0.605	0.582	-
Simplify YOLOv1 [30]	0.542	0.582	0.217	-
Based on YOLOv3 [25]	0.450	0.250	0.321	-
Based on YOLOv3 [16]	-	-	-	0.336

Table 3

Results for the MW-18Mar dataset using the metrics: Precision (P), Recall (R), F1-score (F), and Average Precision (AP).

MW-18Mar	P	R	F	AP
GSAC-DNN (ours)	0.982	0.896	0.937	0.895
GSAC [12]	0.952	0.778	0.856	0.804
Based on YOLOv3 [11]	0.951	0.931	0.941	0.966
AA YOLOv3 [31]	0.939	0.819	0.874	0.884
AB YOLOv3 [31]	0.895	0.902	0.898	0.956
Based on YOLOv3 [25]	0.863	0.759	0.807	0.782
Based on YOLOv3 [16]	-	-	-	0.937

different bounding box size), the one with the highest achieved score is finally selected. Additionally, both detectors have been pre-trained with weights from ImageNet dataset [40], and fine-tuned with the PIROPO database following a transfer learning strategy [41]. Transfer learning consists of re-using pretrained models related to other tasks (for example image classification) for a new task (person detection in our case), assuming that the involved computed feature maps are also (enough) good for other relatively close task domains. Those pretrained models usually are obtained after an intensive training using huge databases (like ImageNet for image classification) giving rise to neural networks models that can represent very efficiently and semantically the visual input information (the images), and therefore they can be re-used for other close related tasks.

Results on the detection performance using the previous databases are provided in Tables 1, 2 and 3, for the metrics P, R, F, and AP. The best results in PIROPO and BOMNI are obtained by GSAC-DNN, while it is the second best in F for MW-18Mar database. Notice that all the detectors use bounding boxes (either upright or rotated), except GSAC-DNN, i.e. they use more prior information (the size and aspect ratio) than GSAC-DNN for computing the pre-

dicted locations. Nonetheless, GSAC-DNN achieves the best scores in two out of three databases and it is very competitive in the other one. Consequently, GSAC-DNN constitutes a more practical solution that can be deployed faster, since creating point-based annotated databases is simply faster. On the other hand, one of the detection systems [16] that achieved the best performance in AP for MW-18Mar, it is significantly worse than GSAC-DNN for the other databases. Consequently, GSAC-DNN not only outperforms [16] in two out of three databases, but is also much more stable in different scenarios. Additionally, the computational cost of GSAC-DNN is the smallest one from all the works of Table 3, as it will be commented later.

The best results after GSAC-DNN for PIROPO correspond to the adaptations of YOLOv3 and Faster-RCNN to work with point-based detections. However, those systems represent more convenient baselines than practical real systems, since multiples models with different fix bounding box sizes must be trained, which, on the one hand, hugely extends the training time; and on the other hand, appropriate bounding box sizes must be found for each different dataset, losing generality capacity. Focusing on the behavior of GSAC-DNN, it is slightly more prone to miss detections than to

Table 4

Results for the three considered datasets using the metrics: False Positives Per Image (FPPI), Miss Rate (MR) and Logarithmic Average Miss Rate (LAMR).

Works	PIROPO			BOMNI			MW-18Mar		
	FPPI	MR	LAMR	FPPI	MR	LAMR	FPPI	MR	LAMR
GSAC-DNN (ours)	0.014	0.075	0.076	0.071	0.077	0.084	0.046	0.104	0.113
GSAC [12]	0.088	0.077	0.183	0.077	0.180	0.337	0.114	0.222	0.341
Based on YOLOv3 [25]	0.350	0.300	0.407	0.400	0.800	0.846	0.200	0.200	0.361
Based on YOLOv3 [16]	-	-	0.392	-	-	0.789	-	-	0.301

Table 5

Ablation experiments for the GSAC-DNN system using the PIROPO database.

Ablation experiments	P	R	F	AP	FPPI	MR	LAMR
GSAC-DNN w/o NMS and w/o conv. classifiers	0.242	0.909	0.383	0.842	2.845	0.090	0.236
GSAC-DNN w/o NMS and with conv. classifiers	0.241	0.925	0.383	0.863	2.909	0.075	0.208
GSAC-DNN with NMS and w/o conv. classifiers	0.989	0.909	0.948	0.906	0.009	0.090	0.090
GSAC-DNN with NMS and with conv. classifiers	0.988	0.925	0.956	0.919	0.014	0.075	0.076

wrong detections, as the Recall and Precision metrics reflect. Although this feature is also shared by most of the other detectors.

Table 4 shows the corresponding results for the metrics FPPI, MR, and LAMR. Like the previous results, the proposed GSAC-DNN outperforms the other works, producing less errors in the detection for the three considered databases. Also, observe that GSAC-DNN achieves a very low rate of FPPI for the three datasets, meaning that a person prediction is very likely to be an actual person. On the other hand, MR is higher than FPPI, supporting the previously mentioned premise that the system is more prone to miss some detections than committing false positives. Nonetheless, this behavior is shared and even is more accused by all the compared detection algorithms.

Regarding the computational cost, only a few of them prove to be able to operate in real time. The fastest one is Adapted YOLOv3, reaching 48 FPS in GPU, that is, 21 ms/image (milliseconds per image) approximately. However, as it was previously commented, it is more a baseline than a real and practical alternative. The next fastest ones are GSAC-DNN with 23 FPS (44 ms/image) and GSAC with 22 FPS (45 ms/image), both in CPU. The others reporting a measurement in GPU are significantly slower: [11] with 7 FPS (143 ms/image), [25] with 6.8 FPS (147 ms/image), Adapted Faster-RCNN with 3 FPS (333 ms/image), AA YOLOv3 with 0.3 FPS (3.3 seconds per image) and AB YOLOv3 with 0.2 FPS (5 seconds per image).

Finally, some ablation experiments for the GSAC-DNN system have been performed using the PIROPO database, which are shown in Table 5. For this purpose, the performance of GSAC-DNN with and without the NMS stage have been evaluated. Also, the convolutional classifiers have been substituted by simple fully-connected classifiers to evaluate the impact of the individual feature extraction step for each classifier. The results evidence the benefits provided by the NMS, which avoids that the number of false positives per image was almost 3. This behavior is expected because the high amount of activations per person allows the system to reach a high R value, and the NMS reduces the false positives, hence, increasing P. About the convolutional classifiers, they propitiate an increase in R, F and AP thanks to the better adaptation of the general feature maps to the necessities of the classifiers.

5. Conclusions

A novel deep neural-network architecture, called GSAC-DNN, is proposed for the task of real-time people detection in omnidirectional images. Unlike other approaches that require large datasets with bounding box annotations (enclosing the person to be detected), GSAC-DNN can be trained with annotations based only on points (each one marking a person). Additionally, it can be globally optimized via an end-to-end training procedure, allowing to improve the performance regarding the standard GSAC. The resulting people detection system, not only has a high accuracy, but also it reduces the cost of annotating the required training databases, allowing that the system can be deployed faster than other detection systems.

CRediT authorship contribution statement

Daniel Fuertes: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing. **Carlos R. del-Blanco:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing. **Pablo Carballeira:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing. **Fernando Jau-reguizar:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing. **Narciso García:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been partially supported by project PID2020-115132RB (SARAOS) funded by MCIN/AEI/10.13039/501100011033 of the Spanish Government.

References

- [1] Y. Shu, Y. Huang, J. Zhang, P. Coué, P. Cheng, J. Chen, K.G. Shin, Gradient-based fingerprinting for indoor localization and tracking, *IEEE Trans. Ind. Electron.* 63 (4) (2016) 2424–2433.
- [2] Z. Zhang, S. He, Y. Shu, Z. Shi, A self-evolving WiFi-based indoor navigation system using smartphones, *IEEE Trans. Mob. Comput.* 19 (8) (2020) 1760–1774.
- [3] S. Venkatesh, R.M. Buehrer, NLOS mitigation using linear programming in ultrawideband location-aware networks, *IEEE Trans. Veh. Technol.* 56 (5) (2007) 3182–3188.
- [4] L. Pei, R. Chen, J. Liu, T. Tenhunen, H. Kuusniemi, Y. Chen, Inquiry-based Bluetooth indoor positioning via RSSI probability distributions, in: 2010 Second International Conference on Advances in Satellite and Space Communications, 2010, pp. 151–156.
- [5] Z. Zhang, Z. Lu, V. Saakian, X. Qin, Q. Chen, L.R. Zheng, Item-level indoor localization with passive UHF RFID based on tag interaction analysis, *IEEE Trans. Ind. Electron.* 61 (4) (2014) 2122–2135.
- [6] F. Li, C. Zhao, G. Ding, J. Gong, C. Liu, F. Zhao, A reliable and accurate indoor localization method using phone inertial sensors, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12, ACM, 2012, pp. 421–430.
- [7] J. Chung, M. Donahoe, C. Schm, I. Kim, P. Razavai, M. Wiseman, Indoor location sensing using geo-magnetism, in: ACM MobiSys, 2011, pp. 141–154.
- [8] Y. Shu, Z. Li, B. Karlsson, Y. Lin, T. Moscibroda, K. Shin, Incrementally-deployable indoor navigation with automatic trace generation, in: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, 2019, pp. 2395–2403.
- [9] D.T. Nguyen, W. Li, P.O. Ogunbona, Human detection from images and videos: a survey, *Pattern Recognit.* 51 (2016) 148–175.
- [10] C. Raghavachari, V. Aparna, S. Chithira, V. Balasubramanian, A comparative study of vision based human detection techniques in people counting applications, *Proc. Comput. Sci.* 58 (2015) 461–469.
- [11] Z. Duan, M. Ozan Tezcan, H. Nakamura, P. Ishwar, J. Konrad, RAPID: rotation-aware people detection in overhead fisheye images, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2020–June, 2020, pp. 2700–2709.
- [12] C.R. del Blanco, P. Carballeira, F. Jaureguizar, N. García, Robust people indoor localization with omnidirectional cameras using a grid of spatial-aware classifiers, *Signal Process. Image Commun.* 93 (2021) 116135, www.gti.ssr.upm.es/data/piropodatabase. (Accessed 8 February 2021).
- [13] A. Bell, T. Mantecón, C. Díaz, C.R. del Blanco, F. Jaureguizar, N. García, A novel system for nighttime vehicle detection based on foveal classifiers with real-time performance, *IEEE Trans. Intell. Transp. Syst.* (2021) 1–13.
- [14] A. Chiang, Y. Wang, Human detection in fish-eye images using hog-based detectors over rotated windows, in: ICME Workshops, IEEE Computer Society, 2014, pp. 1–6.
- [15] R. Seidel, A. Apitzsch, G. Hirtz, Improved person detection on omnidirectional images with non-maxima suppression, *arXiv:1805.08503*, 2019.
- [16] S. Chiang, T. Wang, Y.-F. Chen, Efficient pedestrian detection in top-view fisheye images using compositions of perspective view patches, *Image Vis. Comput.* 105 (2021) 104069.
- [17] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, *arXiv:1804.02767*, 2018.
- [18] R. Seidel, A. Apitzsch, G. Hirtz, Omnidetector: with neural networks to bounding boxes, *arXiv:1805.08503*, 2018.
- [19] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6517–6525.
- [20] H. Kim, E. Chae, G. Jo, J. Paik, Fisheye lens-based surveillance camera for wide field-of-view monitoring, in: 2015 IEEE International Conference on Consumer Electronics, ICCE, 2015, pp. 505–506.
- [21] J. Li, J. Liu, Y. Wang, S. Nishimura, M.S. Kankanalli, Weakly-supervised multi-person action recognition in 360° videos, in: 2020 IEEE Winter Conference on Applications of Computer Vision, WACV, 2020, pp. 497–505.
- [22] O. Krams, N. Kiryati, People detection in top-view fisheye imaging, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS, 2017, pp. 1–6.
- [23] I. Cinaroglu, Y. Bastanlar, A direct approach for human detection with catadioptric omnidirectional cameras, in: 2014 22nd Signal Processing and Communications Applications Conference, SIU, 2014, pp. 2275–2279.
- [24] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2005.
- [25] M. Tamura, S. Horiguchi, T. Murakami, Omnidirectional pedestrian detection by rotation invariant training, in: 2019 IEEE Winter Conference on Applications of Computer Vision, WACV, 2019, pp. 1989–1998.
- [26] G. Cheng, P. Zhou, J. Han, Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images, *IEEE Trans. Geosci. Remote Sens.* 54 (12) (2016) 7405–7415.
- [27] G. Cheng, J. Han, P. Zhou, D. Xu, Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection, *IEEE Trans. Image Process.* 28 (1) (2019) 265–278.
- [28] T. Wang, Y. Hsieh, F. Wong, Y. Chen, Mask-RCNN based people detection using a top-view fisheye camera, in: 2019 International Conference on Technologies and Applications of Artificial Intelligence, TAAI, 2019, pp. 1–4.
- [29] N. Van Tuan, T.B. Nguyen, S. Chung, ConvNets and AGMM based real-time human detection under fisheye camera for embedded surveillance, in: 2016 International Conference on Information and Communication Technology Convergence, ICTC, 2016, pp. 840–845.
- [30] T.B. Nguyen, V.T. Nguyen, S.-T. Chung, S. Cho, Real-time human detection under omnidirectional camera based on CNN with unified detection and AGMM for visual surveillance, *J. Korea Multimed. Soc.* 19 (8) (2016) 1345–1360.
- [31] S. Li, M.O. Tezcan, P. Ishwar, J. Konrad, Supervised people counting using an overhead fisheye camera, in: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS, 2019, pp. 1–8.
- [32] J. Han, D. Zhang, G. Cheng, L. Guo, J. Ren, Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning, *IEEE Trans. Geosci. Remote Sens.* 53 (6) (2015) 3325–3337.
- [33] P.J. Bao, A.I. Maqueda, C.R. Del-Blanco, N. García, Tiny hand gesture recognition without localization via a deep convolutional network, *IEEE Trans. Consum. Electron.* 63 (3) (2017) 251–257.
- [34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.
- [35] B.E. Demiroz, I. Ari, O. Eroglu, A.A. Salah, L. Akarun, Feature-based tracking on a multi-omnidirectional camera dataset, in: International Symposium on Communications, Control and Signal Processing, 2012, pp. 1–5, www.cmpe.boun.edu.tr/pilab/pilabfiles/databases/bomni/. (Accessed 8 February 2021).
- [36] Mirror worlds challenge, www2.icat.vt.edu/mirrorworlds/challenge/index.html. (Accessed 8 February 2021).
- [37] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (VOC) challenge, *Int. J. Comput. Vis.* 88 (2) (2010) 303–338.
- [38] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (4) (2012) 743–761.
- [39] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, vol. 28, Curran Associates, Inc., 2015.
- [40] J. Deng, W. Dong, R. Socher, L. Li, K. Li, Li Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
- [41] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, *arXiv:1808.01974*, 2018.



Daniel Fuertes received the Bachelor of Engineering in Telecommunication Technologies and Services in 2019 and the Master in Signal Theory and Communications in 2020, both from the Universidad Politécnica de Madrid (UPM), Madrid, Spain. Since 2020, he has been a member of the Grupo de Tratamiento de Imágenes (Image Processing Group) of the UPM, where he has been actively involved in several research projects. His research interests include the areas of artificial intelligence, deep learning, computer vision, machine learning, reinforcement learning, and combinatorial optimization.



Carlos R. del-Blanco received the Telecommunication Engineering and Ph.D. degrees in Telecommunication from the Universidad Politécnica de Madrid (UPM), in 2005 and 2011, respectively. Since 2005 he has been a member of the Image Processing Group of the UPM. In addition, since 2011 he is a member of the faculty of the E.T.S. Ingenieros de Telecomunicación, and since 2021 he is Professor of Signal Theory and Communications at the Department of Signals, Systems, and Communications. His professional interests include signal and image processing, computer vision, pattern recognition, machine learning, and stochastic dynamic models. He has been actively involved in European projects and national projects in Spain.



Pablo Carballeira received the Telecommunication Engineering degree (five years engineering program), Communications Technologies and Systems Master degree (two year MS program) and the Ph.D. degree in Telecommunication from the Universidad Politécnica de Madrid (UPM) in 2007, 2010 and 2014 respectively. From 2008 to 2017 he has been a member of the Grupo de Tratamiento de Imágenes (Image Processing Group) at the UPM. Since 2017 he is an

Assistant Professor, and a member of the Video Processing and Understanding Lab, at the Universidad Autónoma de Madrid (UAM). His research interests include computer vision, video coding, and quality of experience evaluation for immersive visual media. He has been actively involved in European projects, national projects, and standardization activities from ISO's Moving Picture Experts Group (MPEG) related to lightfield and free-navigation video technologies.



Fernando Jaureguizar received the degree in telecommunication engineering (six years engineering program) and the Ph.D. degree in telecommunication engineering from the Universidad Politécnica de Madrid (UPM), in 1987 and 1994, respectively. Since 1987, he has been a member of the Image Processing Group, UPM. Since 1991, he has been a member of the faculty of UPM, and since 1995, he has been an Associate Professor of signal theory and commu-

nications with the Department of Signals, Systems, and Communications. His professional interests include digital image processing, video coding,

3DTV, computer vision, and design and development of multimedia communications systems. He has been actively involved in European Projects (Eureka, ACTS, IST, ITEA, and EIT-RM) and national projects in Spain.



Narciso García received the Ingeniero de Telecomunicación degree (five years engineering program) in 1976 (Spanish National Graduation Award) and the Doctor Ingeniero de Telecomunicación degree (PhD in Communications) in 1983 (Doctoral Graduation Award), both from the Universidad Politécnica de Madrid (UPM), Madrid, Spain. Since 1977, he has been a member of the faculty of the UPM, where he is currently a Professor of Signal Theory and Com-

munications. He leads the Grupo de Tratamiento de Imágenes (Image Processing Group), UPM. He has been actively involved in Spanish and European research projects, also serving as an evaluator, a reviewer, an auditor, and an observer of several research and development programs of the European Union. He was a Co-Writer of the EBU proposal, base of the ITU standard for digital transmission of TV at 34–45 Mb/s (ITU-T J.81). He was an Area Coordinator of the Spanish Evaluation Agency (ANEP) from 1990 to 1992 and he was the General Coordinator of the Spanish Commission for the Evaluation of the Research Activity (CNEAI) from 2011 to 2014. He has been the Vice-Rector for International Relations of the Universidad Politécnica de Madrid from 2014 to 2016. He was a recipient of the Junior and Senior Research Awards of the Universidad Politécnica de Madrid in 1987 and 1994, respectively. His current research interests include digital video compression, computer vision, and quality of experience.