Activity Monitoring Made Easier by Smart 360-degree Cameras

Liliana Lo Presti[1][0000−0003−0833−4403], Giuseppe Mazzola[2][0000−0003−3839−9312], and Marco La Cascia[1][0000−0002−8766−6395]

[1] Engineering Department, University of Palermo, Palermo, Italy
liliana.lopresti@unipa.it;
[2] Department of Humanities, University of Palermo, Palermo, Italy
giuseppe.mazzola@unipa.it

Abstract. This paper proposes the use of smart 360-degree cameras for activity monitoring. By exploiting the geometric properties of these cameras and adopting off-the-shelf tracking algorithms adapted to equirectangular images, this paper shows how simple it becomes deploying a camera network, and detecting the presence of pedestrians in predefined regions of interest with minimal information on the camera, namely its height. The paper further shows that smart 360-degree cameras can enhance motion understanding in the environment and proposes a simple method to estimate the heatmap of the scene to highlight regions where pedestrians are more often present. Quantitative and qualitative results demonstrate the effectiveness of the proposed approach.

Keywords: 360 camera; distance estimation; activity understanding

1 Introduction

In recent years, there has been a growing interest in multi-camera systems for activity monitoring. Such systems are often based on distributed smart cameras able to sense the environment, detect pedestrians and objects of interest, and recognize who is doing what.

3D spatial information is especially useful to monitor activities in large, complex areas such as buildings, airports, malls, and crowded environments. To acquire 3D spatial information and use it in multi-camera systems, geometric camera calibration techniques are often used for estimating both intrinsic and extrinsic parameters of the video cameras and the mutual camera positions. While recent progress in the field can make the task at hand easier, still the calibration procedure is time-consuming and requires precise information about the environment, often acquired by imaging predefined patterns.

360-degree camera devices acquire panoramic images with a field of view of 360 and 180 degrees horizontally and vertically, respectively. In the resulting spherical image (stored as equirectangular image), pixels are mapped onto a sphere centered into the camera. It is worth stressing that 360-degree cameras are not ceiling fisheye cameras, since sometimes the two are confused.

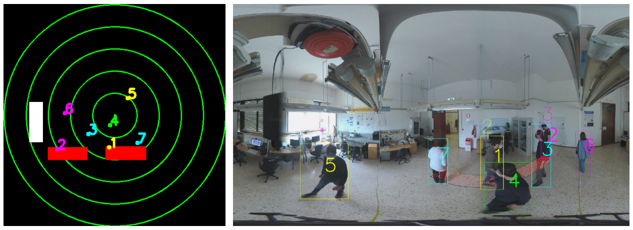


Fig. 1. The image shows a simple scenario where persons move around the environment (ideally a mall or a museum). On the ground, RoIs are marked in black. The goal is to detect when persons are within the RoI given the output of a multi-object tracker. On the left, the polar plot represents the ground plane, the locations of the targets (see ID and colors), and the RoIs (rectangles). The plot is centered on the location of the camera on the ground. Red rectangles are RoIs on which pedestrians have been detected. The same areas are highlighted in red on the equirectangular image shown on the right.

In this paper, we envision a multi-camera system where each smart camera can independently and easily recover spatial information without complex camera calibration procedures. In our system, each smart-camera is a 360-degree. Recently, a new method has been proposed in [16] to estimate the distance of the objects from a 360-degree camera given only its height and the coordinates, in the equirectangular image, of the contact point of the target with the ground plane. The method has been also used in [14] within a tracking technique to estimate the targets' locations onto the ground plane and enhance tracking. Inspired by these former works, we propose:

- a simple method to find correspondences among spherical cameras, thus enabling the deployment of 360-degree camera network;
- a novel method for activity monitoring that uses 360-degree cameras to detect pedestrians within areas of interest with minimal effort;
- a novel method to discover the most visited areas in the scene.

The methods proposed in this paper contribute to show that the use of 360 degrees cameras can simplify activity monitoring by providing spatial information with extremely simple computations and very few information about the environment. Such computation can be easily done onboard of the smart camera by exploiting the geometric properties of the cameras, and adopting an off-the-shelf tracking algorithm adapted to equirectangular images.

To demonstrate our idea, we consider two simple scenarios. In the first, the scene is monitored by two 360-degree cameras and a person is moving around. This scenario is used to demonstrate how to recover correspondences on the ground plane between the two camera predictions. In the second scenario, several persons move in an environment (ideally a mall or a museum) monitored by 360-degree cameras. The goal is detecting when pedestrians are within regions of interest (RoIs) given the output of a multi-object tracker. Figure 1 shows a sample image. For the sake of clarity, RoIs are marked on the ground. Such regions can correspond to areas close to shop windows in a mall or to museum cases. To clarify the task, on the left in Fig. 1 there is a polar plot of the locations of the targets in the monitored area. Rectangles represent RoIs on the ground. Each green circle is one meter apart and rectangles have been drawn considering their (known) locations in the real world, in a reference system centered on the location of the camera on the ground (the projection of the camera on the ground plane). As shown in the figure, red rectangles correspond to RoIs on which pedestrians have been detected on. On the right, the equirectangular image with overlaid the output of the multi-object tracker shows the locations of the pedestrians in the scene.

In Sec. 2, we present related work about activity monitoring techniques based on pedestrians' trajectories. In Sec. 3, we provide details about the geometry of the 360-degree cameras and the relation between equirectangular image pixel coordinates and the ground-plane. In Sec. 4 we present our novel methods while in Sec 5 we report some implementation details. Finally, in Sec. 6 we report experimental results, and in Sec. 7 we discuss conclusions and future work.

2 Related Work

One of the fundamental issue in real-world surveillance is that of optimal camera placement. Furthermore, to get 3D spatial information, camera calibration needs to be performed. Somehow, we are far from systems that can be easily installed or moved because a change in a camera position may require recalibration of the set of deployed cameras.

In [13], coordination between multiple cameras is done through a self-calibration technique using feature correspondences to determine the camera geometry. In particular, planar geometric constraints to moving objects in the scene are used in order to align the scene's ground plane across multiple views. The homography matrix is used to recover the 3D position of the ground plane and the camera positions. This enables them to recover a homography matrix which maps the images to an overhead view. In this paper, we propose the use of smart 360-degree cameras that, in our opinion, represents a step forward towards the building of easily deployable surveillance systems. In our approach, we only need the camera height to find the location on the ground of the pedestrians. Correspondences among 360-degrees cameras (with overlapping field-of-view) can be achieved easily by aligning detections on the ground-plane.

In this paper, we also show how to use 360-degree cameras for simple activity monitoring tasks that are complex to solve with standard cameras. There are several works on vision-based activity monitoring. A simple one is in [10], where activity monitoring is achieved by considering two cameras. First, cameras are calibrated to determine intrinsic and extrinsic parameters by the method in [15]. The calibration involves selecting parallel lines and other features in the ground plane of the image. The result of the calibration is the homography transformation matrix between images coordinates and world coordinates for the camera, as well as the intrinsic and extrinsic (position and orientation) parameters of the camera. The method performs pedestrian tracking and then classifies human motion into classes such as Walking, Stopped, Running, Loitering, Falling and Moving into an area of interest. This transition from image to world coordinates is accomplished by transforming all points measured in the image through the estimated homography matrix. Similarly to [10], we also perform tracking and use the pedestrian location on the ground to detect the kind of activities. In particular, we focus on the standing within areas of interests. In contrast to [10] we do not perform calibration to recover intrinsic and extrinsic camera parameters. Under this point of view, our system is simpler and easy to deploy.

Some earlier works [6, 21, 20, 5] proposed approaches to monitoring activities by using ceiling mounted omnidirectional cameras [5], in particular catadioptric devices [17], that are made of a convex mirror and a camera, pointing up to the mirror. These devices are difficult to calibrate, as the shape of the mirror must be known to compensate the angle-based distortion. We stress here that these devices are different than 360-degree cameras. In [6], a MRF is used to model the background and, hence, detect the foreground. Thus, trajectories on the image plane are recovered by tracking algorithms. No spatial information is recovered and activity is monitored in terms of seconds a person is standing or is walking. In [21], motion detection and people tracking take advantage from motion history images, CamShift and optical flow. A fall detection method for elderly care is also proposed by using a calibrated one-to-one correspondence between the ground locations and the omnidirectional vision sensor images. In [20, 19], the catadioptric sensor is calibrated and used to track moving objects and adjust pan, tilt and zoom parameters of another PTZ camera.

There are many other works focusing on vision-based activity recognition [22, 1, 2, 7, 9, 12]. However, the kind of task these methods consider may largely differ from the one we aim to solve. For the sake of demonstrating that smart 360-degree cameras can enhance activity monitoring we consider the task of detecting if a person is inside/outside an area of interest. We also show that it is easy to analyze the scene and detect the most visited regions.

3 Equirectangular Images and their Geometry

According to the geometrical observations in the work [16], given an equirectangular image, namely the projection of a spherical image acquired by a 360-degree camera, and given the camera height, it is possible to estimate the distance of all the points of the ground plane from their pixel coordinates. For tracking applications, when a target is detected on the image plane, its ground touching point is approximated by the middle point of the lower side of its bounding box. It is then possible to estimate the polar coordinates of the target on the ground in a reference system centered onto the projection of the camera on the ground. In polar coordinates, a target is represented by its distance from the camera and an angle. In this section we provide details on the adopted reference systems.

360-degree cameras are made of at least two lens and can acquire panoramic images with a field of view of 360 degrees horizontally and 180 degrees vertically. Thus, at each shot, it can entirely sense the surrounding environment.

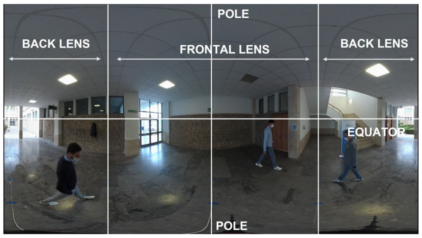


Fig. 2. The figure shows an equirectangular image. The 360-degree camera used to acquire this image is made of two lenses. The left and right image sides have been acquired by one of the two lenses; the central image part has been acquired by the other lens. Thus, the equirectangular image shows how the spherical one is projected. The middle line represents the equator of the spherical image, while the most up and lowest rows are the sphere poles.

In 360-degree cameras, pixels are mapped onto a sphere centered into the camera. Equirectangular and cubic projections are often adopted to make use of these images [8]. In particular, equirectangular images project the whole sphere onto a single image. As shown in Fig. 2, in these images, the central row represents the sphere equator, while the uppermost and lowermost rows correspond to the sphere poles. In general, rows of an equirectangular image correspond to the intersections between the sphere and the planes parallel to the horizontal camera plane [16], and columns are the intersections between the sphere and a vertical plane, including the pole axis, rotated by an angle around the polar axis.

Pixel coordinates (x_r, y_r) of the equirectangular image represent the normalized values of polar and azimuth angles of the corresponding point on the sphere surface. The angles can be recovered from the pixel coordinates by a simple rescaling and shifting such that the polar angle ϕ ranges in $[-90°, 90°]$, while the azimuth angle θ ranges in $[-180°, 180°]$. Of course, by this projection, the radial coordinate of the spherical coordinate system cannot be preserved.

3.1 From equirectangular image to ground-plane coordinates

Fig. 3, on the left, shows the ground plane of the monitored scene and the adopted coordinate reference system. The coordinate reference system is centered on the

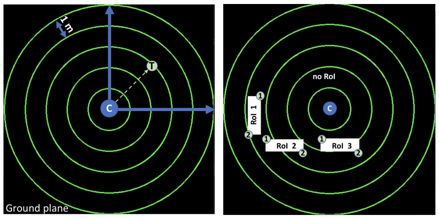


Fig. 3. The image on the left shows the real-world reference system centered on the projection of the camera on the ground. Each circle is one meter apart. The plot represents the overhead view of the ground-plane. On the right, the image shows the locations of the RoIs represented by pairs of points (1 and 2 for each RoI). All points not within a Roi is classified as “no RoI”.

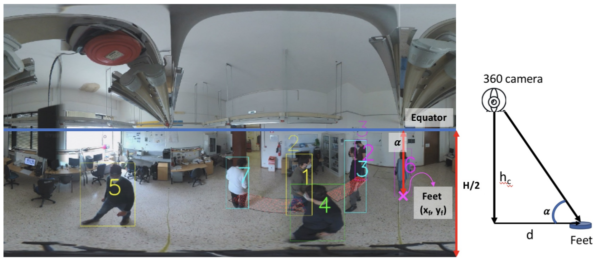


Fig. 4. The image on the left is an equirectangular image overlaid with the tracking output and the equator line. The height of the image is H. For identity 6 (in magenta and to the right of the image), the location of the feet (x_f, y_f) in pixel coordinates is approximated as the middle point of the lower side of the bounding-box enclosing the subject on the image. The angle α is measured as the angular distance from the Equator line (Eq. 2). On the right, in the real world, the distance d on the ground-plane of the subject to the camera can be estimated by Eq. 1 by knowing the camera height h_c and the angle α.

projection of the camera on the ground (C). All circles in the plot are 1 meter apart. Given a point T on the ground plane, we can either consider its Cartesian coordinates as well as its polar coordinates. In the figure, point T has polar coordinates $(3, 45)$ namely the distance on the ground from the camera is 3 meters and the polar angle is 45 degrees.

Assuming the horizontal camera plane is parallel to the ground-plane (i.e., the camera roll angle is equal to 0), the only information needed to associate equirectangular image pixel coordinates to a ground-plane point in the real world is the camera height h_c. As reported in [16] and shown in Fig. 4, given a point on the ground, the distance d from the camera can be estimated as:

$$d = h_c \cot \alpha \tag{1}$$

where α is the angle between the ground plane and the line through the camera center and the point on the ground. Fig. 4 aims at representing the angle α given the target bounding-box. Let us consider the subject at the extreme right (in magenta color). The point on the ground in pixel coordinates is approximated by the middle point P of the lower side of the bounding-box. Let us assume that the height of the equirectangular image is H and that $P = (x_f, y_f)$. Thus, by applying the equation

$$\alpha = \frac{\frac{H}{2} - y_f}{\frac{H}{2}} \cdot 90^\circ \tag{2}$$

we can recover the angle α. The coordinate x_f is normalized and used to find the azimuth angle θ. Therefore we can transform each point on the ground from the equirectangular pixel coordinates (x_f, y_f) to the corresponding polar coordinates in real-world (d, θ).

4 Activity Monitoring in 360-degree camera network

Activity monitoring is a wide field in computer vision. Many tasks require that the location on the ground plane of the targets is known and camera calibration techniques to estimate intrinsic and extrinsic camera parameters are applied.

As already explained in Sec. 3, location on the ground plane can be recovered by each 360-degree camera independently. Here we show that, when the exact relative camera pose (horizontal orientation and relative position) is unknown, correspondences between the reference systems of the cameras can be found by a simple alignment technique. We also focus on two applications. In the first one, we aim at detecting whether a target stands within an apriori known area of interest. The second application refers to the detection of the areas mostly frequented in the monitored environment.

4.1 Correspondences between the camera reference systems

We consider a system of two 360-degree cameras. Fig. 6 shows the trajectory on the ground-plane of the same pedestrian as estimated independently by the two cameras. Despite the ground plane is the same, the points are represented in different coordinate systems. Hence, we need to find the geometrical transformation that allows changing the coordinate reference system. The problem can also be seen as a point clouds alignment problem and is solved by working in Cartesian coordinates and establishing the correspondences between points (x_1, y_1) and (x_2, y_2) in the two reference systems respectively. The correspondence is modeled as a roto-translation transformation:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} cos\psi & -sin\psi \\ sin\psi & cos\psi \end{bmatrix} \cdot \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} + \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} \tag{3}$$

where ψ is the rotation angle while $[\delta x, \delta y]$ represent the translation coefficients.

In our method, we simply use the trajectories estimated independently by each camera on the ground-plane by using a pedestrian tracker adapted to equirectangular images. Thus, we consider those frames where the person is detected in both the cameras and estimate the geometrical transformation that aligns the points. Such estimation can be done by minimizing the mean square error using the least-squares (LS) method. The method works if the horizontal camera plane is parallel to the ground-plane. When this assumption does not hold, the roll angle of each camera must be known, and equirectangular images need to be corrected to account for it.

4.2 Detecting activities within areas of interests

One challenging task in vision-based activity monitoring is the detection of pedestrians moving into areas of interest. It is an important task in several applications such as: surveillance (pedestrians enter a restricted access area), cultural heritage (visitors are too close to an art opera) or retail applications (customers are more interested to a shop rather than another). In all these fields, areas of interest are generally apriori known and defined. The complex task is to recover, from the visual information, the pedestrians' locations on the ground to estimate their position with respect to the area of interest (inside or outside). Here, we show that the task becomes extremely simple when 360-degree cameras are used.

In our application, RoIs are modeled by means of their actual real-world coordinates. In Fig. 3, on the right, three regions of interest are considered. Our method does require two points on the ground to describe each RoI.

To detect if a pedestrian is within an area of interest, we model his/her location on the ground by means of the middle point of the lower side of the bounding box estimated on the equirectangular image. Then, by means of equations 1 and 2 we estimate the location on the ground plane in real world coordinates and compare it with the RoI position in the scene. No further processing is required.

4.3 Detecting areas of interest in the scene

When areas of interest are not provided, or it is of interest to detect what part of the scene has been the most or the least frequented by the pedestrians, it is useful to build a discretized heatmap of the environment. In our approach, it is possible to collect accurate measurements of the pedestrian locations on the ground-plane and the estimation of the heatmap is straightforward.

As shown in Fig. 5, it is sufficient to divide the ground plane into circular bins. In particular, we divided it into N circular sectors, with a fixed step angle (in the image on the left, 45 degrees, namely $N = 8$). Then each circular sector is divided into $M+1$ circular bins. Such circular bins can be easily re-projected onto the equirectangular image. Based on the equirectangular projection properties, circles in the real world are mapped into lines, and circular sectors are mapped into vertical stripes (see the image on the right in Fig. 5). Thus, each circular bin is a rectangle in the equirectangular image. When the distance from the camera increases, the height of the rectangles decreases, and, on the equirectangular image, the bins are not uniformly distributed in the vertical direction.

The heatmap can be easily computed by incrementing the circular bins each time a pedestrian is located inside the cell represented by the bin itself. The computation can be carried on both in polar and in pixel coordinates given a precomputed grid on the equirectangular image. Once the heatmap is computed, it is possible to recover the most trampled area in the scene.

5 Implementation Details

All proposed methods rely on the output of a multi-object tracking algorithm. Any MOT method can be adapted to the tracking in equirectangular images, provided that circularity of the image is taken into account. Despite tracking is not the focus of this paper, here we describe the strategy adopted to get the pedestrians' bounding-boxes.

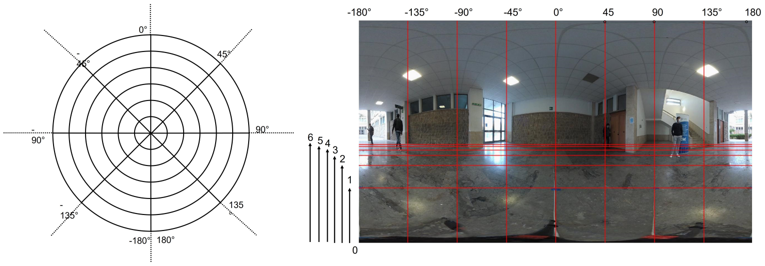


Fig. 5. On the left, the image shows how the ground plane is partitioned in circular bins to accumulate information about the most visited sites in the scene. The image on the right shows how the binarization of the polar space is remapped onto the equirectangular image.

Tracking on the ground plane by 360-degree cameras has already been proposed in [14]. The work describes a simple MOT strategy, based on the tracking-by-detection paradigm, that uses Faster-RCNN [18] to locate persons on the image plane. To account for the image circularity, the image is expanded at both the sides and duplicated detected bounding-boxes are removed.

The target's location on the ground is used in [14] to enhance the association between new detections and predicted target locations. The latter are computed by using the Kalman filter while data association is solved by the Munkres algorithm. The data association matrix combines the distance among targets in the real world and appearance features extracted from a ResNet-50 [11] model. Furthermore, thresholds are used to decide when to add a new identity to the target pool and when to kill an identity by removing it from the same pool.

In this paper, we implemented our own version of the work in [14] with some small changes. First, we detect pedestrians by YOLOv5 [4], which provides better detection and has a lower missing rate. Considering that the detections are provided at a good rate also in case of severe partial occlusions, we avoided using the Kalman filter. We simplified the data association strategy by considering an approach similar to that used in SORT [3]. First, we associate the most recently detected targets, and later the ones missing from the scene for more time. We also store the appearance features of the last 20 frames and use them to compute the association matrix. Since in crowded environments occlusions are more frequent, we found it preferable to compute smaller tracks than to risk higher identity switch rates, and therefore we decreased the threshold to kill missing identities. Then, we applied a post-processing step to rejoin the tracks based on similarity in position and appearance.

6 Experiments

There are not many publicly available dataset of videos acquired by 360-degree cameras, and none of the public ones is multi-camera nor focuses on activity monitoring. Therefore, we collected and manually annotated videos to demon-

Table 1. Camera Correspondences Evaluation

Method	ψ (deg)	$\delta x(m)$	$\delta(m)y$	rmse (m)
LS-M	2.29	2.90	0.16	0.20
true values	2.00	3.00	0.00	–

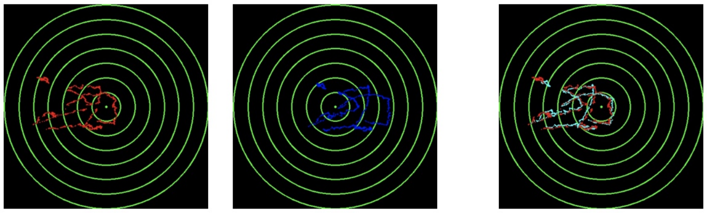
strate the effectiveness of the proposed methods as described in the following.

6.1 Finding correspondences between camera reference systems

We acquired two videos by using two 360-degree cameras. In this scenario, only two persons were moving around the scene. We used the trajectories estimated from the two camera videos of one person to estimate the roto-translation matrix, and tested the correspondences on the trajectories of the second person. Each trajectory includes around 1800 points.

Table 1 reports the estimated values of the rotation angle ψ, and the translation coefficients (in meters) obtained from the training trajectories, and the root mean squared error on the test trajectories. We also report the manually estimated values.

Comparing the estimated values to the true ones, there is an error lower than half degree in the estimated angle ψ and of few centimeters in the translation coefficients (10 and 16 in the two directions respectively). The rmse is of around 20 cm. From the analysis of the results we concluded that the parameter estimation is affected by the precision by which feet are approximated on the ground. The approximation in turn depends on the quality of the detection (which may not be accurate, especially in case of partial occlusions). Moreover, while a camera can get a frontal view of a person, the other camera can get a side view. In these cases, the estimated feet location on the ground may refer to different 3D points. Another issue we noticed is that the horizontal plane of one of the two cameras was not perfectly parallel to the ground plane. This explains the measured error on the translation coefficients and is also visible in Fig. 6. The figure shows the test trajectory of the pedestrian in the reference system of camera 1 and camera 2. On the right, the figure shows the aligned trajectories. As shown in the figure, the error becomes more evident 3-4 meters far from the camera.



Trajectory - Camera 1

Trajectory - Camera 2

Aligned Trajectories in Camera 1

Fig. 6. On the left, the trajectories of the test person in the reference systems of camera 1 and 2 respectively. On the right, the trajectory detected by camera 2 is re-projected in the reference system of camera 1 by the estimated roto-translation matrix.

Table 2. Confusion matrix: actual (rows) vs. predicted classes (columns).

Class	no RoI	RoI 1	RoI 2	RoI 3
no RoI	**0.982**	0.007	0.005	0.005
RoI 1	0.182	**0.810**	0.008	0
RoI 2	0.185	0	**0.815**	0
RoI 3	0.226	0	0.001	**0.773**

Table 3. Precision, recall and F1-Score for each class.

Class	Precision	Recall	F1-Score
no RoI	0.913	0.982	0.947
RoI 1	0.940	0.810	0.870
RoI 2	0.955	0.815	0.879
RoI 3	0.959	0.773	0.856

6.2 Activity Detection within RoI

To assess the ability of our approach in detecting activities within RoIs, we collected and manually annotated two videos of persons moving in the scene. The first video counts 2400 frames, with 13 persons entering/exiting the scene. The second video counts 2436 and 7 persons moving around. Fig. 1 represents one sample image from the first video. We marked on the ground of the scene three RoIs to facilitate manual annotation of the data. Only the identity of each person and the RoI on which the person is on have been manually annotated. The person is considered inside an RoI when he/she stands inside the rectangular area or on the RoI boundaries with both feet. The person is outside the RoI when one of the feet is outside the RoI. Fig. 7 shows samples of images and corresponding annotations. Overall, the videos used to test our technique are very challenging due to the frequent occlusions and changes of directions of the pedestrians. The subjects involved in the experiments did not know anything about the method we wanted to test. It has been explained to them that the rectangles on the ground were marking the area close to shop windows and asked them to simulate a visit to a mall.

To assess the capability of our method to infer the presence of a person within an RoI, we treat it as a multi-class classification problem.

At each frame and to each target, we assign a label indicating that the person is not inside an RoI or is inside one of the three RoIs. Thus, there are overall 4 classes. We compare the estimates of our method against the ground-truth and computed the confusion matrix, shown in Table 2, and metrics such as precision, recall, and F1-score, reported in Table 3.

Of course, persons stand inside the RoIs for a time that is lower with respect to the time they spend outside and far from the RoIs (in areas where confusion among the two kinds of classes is not possible). Thus, recall of the class "no RoI"

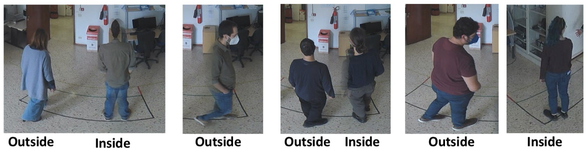


Fig. 7. The figure shows samples and corresponding annotation label. When feet are visible within the RoI or on its boundary, the assigned label is inside the RoI, otherwise the subject is considered outside of the RoI (class "no RoI").

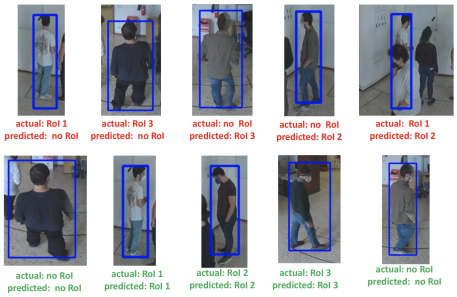


Fig. 8. The top row of the image shows samples on which our method fails to predict the correct class. The bottom row shows success cases of our approach.

is especially high. As shown in Table 2, there is very little confusion among the RoI classes. Only between adjacent RoIs there might be some confusion, when a person moves fast from an RoI to the adjacent one. As expected, most of the confusion is between the class “no RoI” and the RoI classes. The confusion is especially related to the exact time when a person enters the RoI. In some cases, our method detects earlier when a person is entering the scene, in other cases the method needs to wait for a few frames before estimating that a person is outside the area of interest. The main reason why this happens is that the feet of the pedestrians are approximated as the middle point of the bounding-boxes and there is not a precise feet localization. Also, the trackers may provide incorrect bounding-boxes during occlusions.

As shown in Table 3, all metrics for the RoI classes are comparable, which indicates that the performance of the method is independent of the location of the RoI with respect to the camera. Of course, RoI must be at a reasonable distance from the camera. In our experiments, the RoIs were at no more than 2, 3, and 4 meters of distance (see Fig. 3). In other experiments we have seen that our method works fine at distances lower than 6 meters. Beyond this distance, the method may be inaccurate mostly because the pedestrian detector is not able to accurately locate people on the equirectangular image.

Overall, these experiments show the viability of the approach. Some cases of success and failures of the method are shown in Fig. 8. As the images confirm, some failure cases are ascribable to inaccurate detection or occlusions.

6.3 Detecting Areas of Interest

The same two videos used to test the activity detection within RoIs were used to estimate an heatmap of the environment. We also estimate heatmaps on the CVIP360 dataset [16] including 11 indoor videos, and 6 outdoor videos.

To detect areas of interest, we used the location of the pedestrians on the ground and computed the discretized heatmap presented in Sec. 4.3. This is somewhat the inverse problem where, given the pedestrians' locations, we aim at discovering the most visited areas in the scene.

Fig. 9 shows the map of the room and the heatmap estimated from the tracking results on the two collected videos. In the map, circles are $0.5m$ apart.

Since persons walk but also stands at specific points in the scene, it is possible to visually discover areas (in red) that have been more frequented. Comparing the polar coordinates of these regions with those in Fig. 3, it is possible to state that these regions correspond to our RoIs.

The CVIP360 dataset includes the manual annotations of the bounding boxes of the people in the equirectangular image, which we used to estimate the pedestrians' location on the ground. We computed two cumulative heat-maps, shown in Fig. 10, one for the videos taken outdoor and the other for the videos taken indoor. The maps show the density of the most trampled areas on the ground plane. Circles are $1m$ apart. Note that the heatmaps have $N \times (M + 1)$ bins, where N is the number of circular sectors (8 in our experiments) and M is the number of circular crowns, which depends on the maximum annotated distance on the ground plane ($M = 6$ for the indoor videos, $M = 10$ for the outdoor ones) and on the quantization step (1 meter, in our experiments). With respect to the map in Fig. 9, no evident areas of interest is highlighted. In fact, the CVIP360 dataset is meant for tracking purposes, and pedestrians continuously move around the scene. This is very evident in the heatmaps where circular bins have comparable values. Thus this experiment confirm the viability of the proposed approach.

7 Conclusions and Future Work

This paper shows how smart 360-degree cameras can enhance multi-camera surveillance systems in several respects. Firstly, with one 360-degree camera at the center of the scene, it is possible to sense the surrounding environment. This choice limits blind spots, common in standard multi-camera systems. Secondly, smart 360-degree cameras and, more in general, equirectangular image processing do not need complex calibration techniques to recover the locations of the objects on the ground plane. The only information required is the camera height. This simplifies the deployment of the cameras, which does not require specialized technical skills. Thirdly, as shown in this paper, smart 360-degree cameras enhance and simplify some activity monitoring tasks, such as detecting

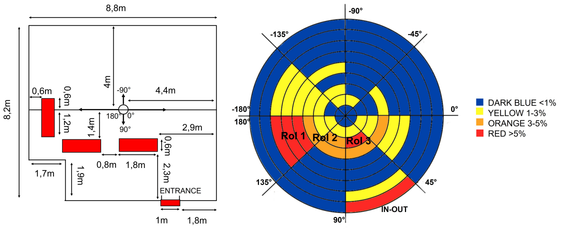


Fig. 9. On the left, the planimetry of the room where we conducted our experiments. RoIs are colored in red. The optical axis of the camera was parallel to the longer side and the front lens pointed to the right. On the right, the estimated heatmap with circles 0.5 meter apart. Red and orange areas are related to the RoIs, and the entrance.

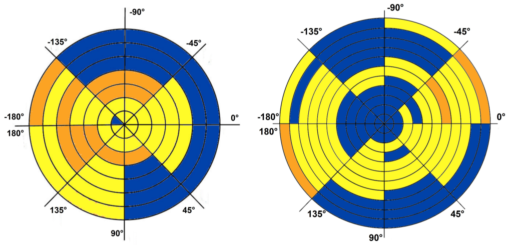


Fig. 10. The two cumulative heatmaps of the indoor(left) and outdoor(right) videos of the CVIP360 dataset. Circles are spaced 1 meter apart. Color notation is the same of Fig. 9. No evident areas of interest emerge.

the standing within areas of interest. The computation required to estimate the relative location of pedestrian and regions of interest are so simple that can be carried on the devices without increasing the computational complexity of the algorithms on the smart camera. Finally, this paper shows that projections on the ground plane of the targets' locations in different 360-degree cameras are related by simple roto-translation transformation that can be easily estimated by state-of-the-art techniques.

There are also some limitations to consider that will be the core of future investigations. When mounting the camera, there might be some small roll angle to account for when estimating correspondences between the camera ground-plane projections. We will study how to include the automatic estimation of this parameter to find better correspondences between the camera reference systems.

360-degree cameras have a blind spot exactly at their bottom. Furthermore, distortion near the poles is so large that traditional pedestrian detectors are unable to deal with it. Specialized detectors are needed to deal with such cases. Our experiments have shown that the method is very sensitive to the accuracy by which feet are detected, especially moving far from the camera. On one hand, it requires improving feet localization on the image. On the other hand, it requires a better strategy to deal with partial occlusions that make it impossible to estimate the location on the ground plane.

Despite these limitations, 360° cameras are very appealing for practical multi-camera system applications and their use could spread quickly in the near future.

8 Acknowledgement

This work was partially supported by the Italian MIUR within PRIN 2017, Project Grant 20172BH297_004: I-MALL - improving the customer experience in stores by intelligent computer vision.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. Acm Computing Surveys (Csur) **43**(3), 1–43 (2011)

2. Beddiar, D.R., Nini, B., Sabokrou, M., Hadid, A.: Vision-based human activity recognition: a survey. Multimedia Tools and Applications **79**(41), 30509–30555 (2020)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016)
4. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
5. Boult, T., Qian, C., Yin, W., Erkin, A., Lewis, P., Power, C., Micheals, R.: Applications of omnidirectional imaging: Multi-body tracking and remote reality. In: Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV'98 (Cat. No. 98EX201). pp. 242–243. IEEE (1998)
6. Chen, X., Yang, J.: Towards monitoring human activities using an omnidirectional camera. In: Proceedings. Fourth IEEE International Conference on Multimodal Interfaces. pp. 423–428. IEEE (2002)
7. Climent-Pérez, P., Spinsante, S., Mihailidis, A., Florez-Revuelta, F.: A review on video-based active and assisted living technologies for automated lifelogging. Expert Systems with Applications **139**, 112847 (2020)
8. Corbillon, X., Simon, G., Devlic, A., Chakareski, J.: Viewport-adaptive navigable 360-degree video delivery. In: 2017 IEEE international conference on communications (ICC). pp. 1–7. IEEE (2017)
9. Demiröz, B.E., Ari, I., Eroğlu, O., Salah, A.A., Akarun, L.: Feature-based tracking on a multi-omnidirectional camera dataset. In: 2012 5th International Symposium on Communications, Control and Signal Processing. pp. 1–5. IEEE (2012)
10. Fiore, L., Fehr, D., Bodor, R., Drenner, A., Somasundaram, G., Papanikolopoulos, N.: Multi-camera human activity monitoring. Journal of Intelligent and Robotic Systems **52**(1), 5–43 (2008)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Kobilarov, M., Sukhatme, G., Hyams, J., Batavia, P.: People tracking and following with mobile robot using an omnidirectional camera and a laser. In: Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006. pp. 557–562. IEEE (2006)
13. Lee, L., Romano, R., Stein, G.: Monitoring activities from multiple video streams: Establishing a common coordinate frame. IEEE Transactions on pattern analysis and machine intelligence **22**(8), 758–767 (2000)
14. Lo Presti, L., Mazzola, G., Averna, G., Ardizzone, E., La Cascia, M.: Depth-aware multi-object tracking in spherical videos. In: International Conference on Image Analysis and Processing. pp. 362–374. Springer (2022)
15. Masoud, O., Papanikolopoulos, N.P.: Using geometric primitives to calibrate traffic scenes. Transportation Research Part C: Emerging Technologies **15**(6), 361–379 (2007)
16. Mazzola, G., Lo Presti, L., Ardizzone, E., La Cascia, M.: A dataset of annotated omnidirectional videos for distancing applications. Journal of Imaging **7**(8), 158 (2021)
17. Nayar, S.K.: Catadioptric omnidirectional camera. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition. pp. 482–488. IEEE (1997)

18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
19. Scotti, G., Marcenaro, L., Coelho, C., Selvaggi, F., Regazzoni, C.: A novel dual camera intelligent sensor for high definition 360 degrees surveillance. Intelligent Distributed Surveilliance Systems pp. 26–30 (2004). https://doi.org/10.1049/ic:20040093
20. Scotti, G., Marcenaro, L., Coelho, C., Selvaggi, F., Regazzoni, C.: Dual camera intelligent sensor for high definition 360 degrees surveillance. IEE Proceedings-Vision, Image and Signal Processing **152**(2), 250–257 (2005)
21. Wang, M.L., Huang, C.C., Lin, H.Y.: An intelligent surveillance system based on an omnidirectional vision sensor. In: 2006 IEEE Conference on Cybernetics and Intelligent Systems. pp. 1–6. IEEE (2006)
22. Zhou, Z., Chen, X., Chung, Y.C., He, Z., Han, T.X., Keller, J.M.: Activity analysis, summarization, and visualization for indoor human activity monitoring. IEEE transactions on circuits and systems for video technology **18**(11), 1489–1498 (2008)