Inter-Homines: Distance-Based Risk Estimation

Matteo Fabbri, Fabio Lanzi, Riccardo Gasparini, Simone Calderara, Lorenzo Baraldi and Rita Cucchiara
Department of Engineering “Enzo Ferrari”
University of Modena and Reggio Emilia
Email: name.surname@unimore.it

***Abstract*—In this document, we report our proposal for modeling the risk of possible contagiousity in a given area monitored by RGB cameras where people freely move and interact. Our system, called Inter-Homines, evaluates in real-time the contagion risk in a monitored area by analyzing video streams: it is able to locate people in 3D space, calculate interpersonal distances and predict risk levels by building dynamic maps of the monitored area. Inter-Homines works both indoor and outdoor, in public and private crowded areas. The software is applicable to already installed cameras or low-cost cameras on industrial PCs, equipped with an additional embedded edge-AI system for temporary measurements.**

I. Risk Model

Let’s consider a scene with N people $k_0, \dots, k_{N-1}$ at a given time t. Given two subjects k_i and k_j with distance $d_{i,j}$, we define their reciprocal risk as follows:

$$rr_{i,j}^{(t)} = \eta \, e^{-\beta \max(0,\, d_{i,j} - \tau)} \tag{1}$$

where η, β and τ are parameters that respectively control height, slope and the full width at maximum of the function. In this specific application, η is a mitigator used to decrease the risk when some criteria are met, e.g., when at least one of the two people is wearing a facial mask. β, instead, controls how the risk decreases when the distance is greater than τ and can model environmental characteristics such as air temperature. Lastly, τ, controls the transmissibility of the disease via respiratory droplets and define the minimal distance allowed between two people. It should follow World Health Organization and national guidelines but can be further increased to better preserve the safety of people in critical places such as COVID-19 hospital units. We then define the individual risk at time t as:

$$R_i^{(t)} = \max_{j=0 \dots N-1, j \neq i} \{rr_{i,j}^{(t)}\} \tag{2}$$

The global risk at t of the scene is then computed as follows:

$$G^{(t)} = \min\Big(1,\, \frac{1}{C} \sum_{i=0}^{N-1} R_i^{(t)}\Big) \tag{3}$$

where C is the maximum capacity of the scene. This capacity can be either given by the user or calculated using covering algorithms. Finally. the dynamic global risk is computed as:

$$D^{(t)} = \frac{1}{W} \sum_{w=0}^{W-1} G^{(t-w)} \tag{4}$$

where W is the size of the temporal window. At a given time t, $D^{(t)} \in [0, 1]$ is the global risk of the scene and it is used to trigger alarms when it reaches a given threshold.

II. Inter-Homines Technical Core

Here we give an overview of the pipeline we used to process videos in real time. The aim of our Inter-Homines system is to detect people, compute their distance and provide a dynamic risk level of the area, as well as producing a human readable visualization with anonymized people. For GDPR constraints, no visual data is recorded but, instead, only people coordinates are extracted and stored. The following subsections summarize the key elements of our system.

A. People Detection

As we are interested in the best speed-accuracy trade-of, we choose CenterNet [1] as detector. In particular, we rely on the DLA backbone [2] which yields 51.3% AP for the people class on MS COCO [3], running at 52 FPS on a Titian XP.

CenterNet is capable of producing a precise localization of every person in the image, however, it does not take into account occlusions that usually happen in real world scenarios. If a person is occluded by an object or by other people, CenterNet predicts a tight bounding box that only contains the visible part of the person, ignoring his full shape. This usually happens with the bottom part of the body, as the camera is commonly placed several meters above the ground. Since we are interested in recovering the ground plane coordinate of each person through homograpy, we need to know the position (in image plane) of the feet of each detected person. This task cannot be accomplished by solely relying on CenterNet.

B. Feet and Head Localization

To overcome the aforementioned limitations without introducing complexity to the overall system, we propose to utilize a small network to predict the feet position given a bounding box containing a person, even if the feet are not visible.

To this aim we rely on a simple but effective CNN that, given an image M tightly containing a person, it regresses to the midpoint $P_f = (x_f, y_f)$ of the segment having the two feet as endpoints. This ensures that we know the exact position in image plane where every person touches the ground. Since we are also interested in anonymizing the face of each detected person, we further predict the location of the head $P_h = (x_h, y_h)$. To that purpose we utilize Resnet50 [4] as backbone. Training has been conducted on JTA [5].

This step ensures a more precise localization of the feet while also coping with truncated bounding boxes. Our network can effectively obtain an accurate position of each head and it is used to extend the bounding box to its regular shape.

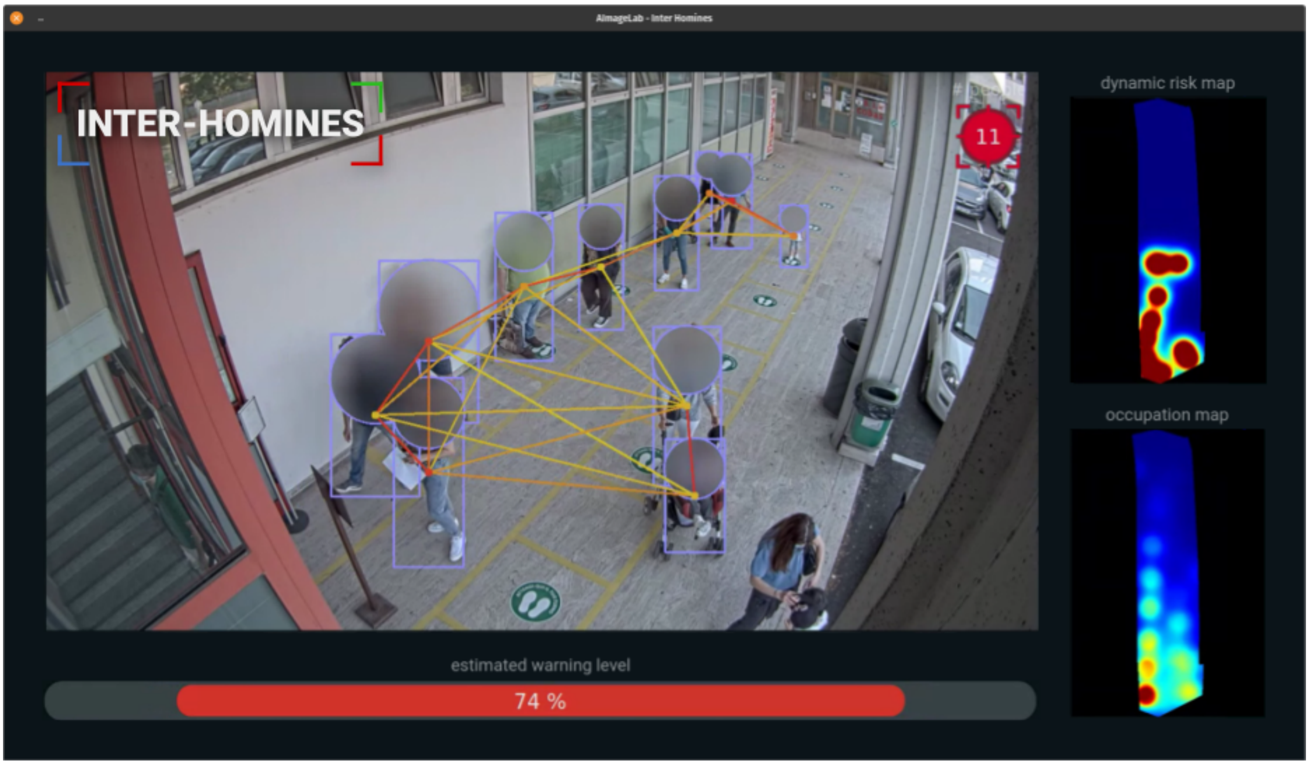

Fig. 1. GUI of our system. In the main frame, anonymized bounding boxes are superimposed to the image. Colored links encodes people reciprocal distance. On the right, two maps shows the bird-eye view of the area. The estimated risk level of the scene resides at the bottom of the interface.

C. *From Image Plane to Ground Plane*

In order to recover the ground plane coordinates given image plane coordinates, we relied on an homography transformation. A practical way to calculate the homography matrix H is to find a set of at least four points pairs of target and source planes and to minimize the back-projection error.

To easily obtain the points pairs of image and ground planes we designed a simple procedure that consists in placing nine markers at the center of the monitored area. By means of a simple graphic interface, the user can take a snapshot of the camera and click with the cursor the centers of the nine markers in order to acquire the pixel coordinates. Once the nine pairs of points have been identified and the homography matrix calculated, the carpet can be safely removed and the system will continue to work properly as long as the camera maintains its position.

Now that we have the 3D position of every person in the scene, the dynamic global risk in Eq. 4 can be calculated and given as output along with other information that we summarize in the following subsection.

D. *System Output*

A convenient graphical interface highlights all the main results of the analysis of our Inter-Homines system, allowing to evaluate at a glance the crowding conditions in the monitored area (see Fig. 1). This interface is made with Qt to guarantee compatibility with all the main operating systems.

a) Anonymized Frame: It shows real-time the bounding box detections superimposed to the input RGB frame. The system is privacy compliant and all the faces are obscured. The colors of the segments that links people indicate the extent of the infraction, going from a dark red for the most serious infraction to yellow for the minor ones.

b) People Counter: At the top right of the frame we also display the number of detected people updated in real time. This number is an average computed in a window of W frames to account for miss detections and false positives.

c) Dynamic Risk and Occupation Maps: In the right part of the interface two bird eye views of the walking area are updated real-time. The Dynamic Risk map shows a snapshot of the current situation of the area. The Occupation map, instead, displays the overall aggregated risk and it is computed by averaging the Dynamic Risk maps of the whole day.

d) Estimated Warning Level: In the lower part of the window a bar represents the estimated risk in the monitored area and it is computed using Eq. 4. The application provides the possibility to send an alarm signal if certain thresholds on the number of people or on the risk level are exceeded.

e) Weekly Report: Since we want to give insightful statistics to help with the prevention of the infection, our system periodically produce a report. The report contains statistics about number of people, risk level, number of infractions and occupation maps aggregated by hours and days. To this end we utilize a non-relational database to store timestamp and position of each person captured by our system.

References

[1] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv:1904.07850*, 2019.

[2] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[5] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara, "Learning to detect and track visible and occluded body joints in a virtual world," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.