ACTA IMEKO
ISSN: 2221-870X
June 2023, Volume 12, Number 2, 1 - 11

Monte Carlo human identification refinement using joints uncertainty

Mariolino De Cecco[1], Alessandro Luchetti[1], Mattia Tavernini[2]

[1] *Dep. of Industrial Engineering, University of Trento, Via Sommarive 9 - 38123 Trento (TN), Italy*
[2] *Robosense Srl, Viale Dante, 300 - 38057 Pergine Valsugana (TN), Italy*

ABSTRACT
In this work, we propose a new method to re-identify the same individual among different people using RGB-D data.
Each human signature is a combination of soft biometric traits. In particular, we extract a color-based descriptor and a local feature descriptor through a Monte Carlo-based algorithm taking into account the uncertainty of human joints and, applied to each descriptor, refines the similarity match against a spatiotemporal database that updates over time.
We analyzed the effects of Monte Carlo refinement in terms of the final maximum matching score obtained for the two descriptors. In addition, we tested the performance of the proposed method on a widely used public dataset against one of the best re-identification methods in the literature. Our method achieves an average recognition rate of 99.1 % rank-1 without identification error.
Its robustness also makes it suitable for industrial applications.

Section: RESEARCH PAPER

Keywords: human re-identification; uncertainty of human joints; Monte Carlo method; RGB-D camera

Citation: Alessandro Luchetti, Mattia Tavernini, Mariolino De Cecco, Monte Carlo human identification refinement using joints uncertainty, Acta IMEKO, vol. 12, no. 2, article 32, June 2023, identifier: IMEKO-ACTA-12 (2023)-02-32

Section Editor: Alfredo Cigada, Politecnico di Milano, Italy, Roberto Montanini, Università degli Studi di Messina, Italy

Received December 3, 2022; **In final form** February 21, 2023; **Published** June 2023

Corresponding author: Alessandro Luchetti, e-mail: alessandro.luchetti@unitn.it

1. INTRODUCTION

Nowadays, one challenge still open is the automatic re-identification (RE-ID) of people, which involves re-identifying the same person in environments populated by multiple people. It consists of robustly re-identifying a person even after changes such as occlusions, lighting variability, different backgrounds, and human poses. Most works in this field involve mass surveillance applications [1]. The widespread use of surveillance cameras in public places such as streets, malls, airports, or large-scale events makes RE-ID of people mainly worthwhile to increase public safety standards [2]. Other applications range from long-term pedestrian tracking [3] to human activity recognition [4]. For each of them, automatic classification systems can identify and differentiate people based on many pre-recorded or real-time videos with low human effort and time.

At the same time, however, the need to have a robust choice of the identified person is essential in the industrial environment, where autonomous mobile robots operate in environments populated by humans and interact with them in several ways [5], [6]. The goal is to preserve the safety of the operators while supporting their work: proper identification (ID) of operators plays a key role in this context.

In this study we explored and selected new descriptors and methods to address the task of RE-ID. In each frame, we combined color and 3D data acquired by an RGB-D sensor to extract color, biomechanics, and depth information. We designed a new method to select and track a person with high accuracy suitable for industrial applications, considering the uncertainty of human joints. To demonstrate the performance of the proposed method, we first evaluated the Monte Carlo refinement effects and then the whole method in terms of recognition rate on a public BIWI RGBD-ID dataset. Furthermore, we compared the obtained results with a public RE-ID neural network provided by the OpenVINO toolkit [7] on the same dataset.

The rest of the paper is organized as follows. Section 2 outlines the state of the art in RE-ID techniques. In Section 3, we describe the method developed. In section 4, we discuss the effects of Monte Carlo refinement and validation. In the final section, we draw conclusions.

2. RELATED WORK

Most of the techniques used in literature for RE-ID exploit soft biometric traits to distinguish humans from their peers. Soft biometric traits [8] are physical (e.g., skin color, eye color, hair color, height, weight, gender, race, etc.), behavioral (e.g., gait, keystroke, signature, etc.), or adhered human characteristics (e.g., clothes color, tattoos, accessories, etc.). Their advantages include compliance with natural human description labels and non-obtrusiveness in data acquisition. In general, colour is the most common factor affecting the RE-ID performance and is often encoded in histograms [9], [10]. Its inconsistency is the most prominent with respect to viewpoints and poses changes [11]. For these reasons, we also used a color-based descriptor in our method. Other approaches use matching based on local feature detectors and descriptors, such as Speed Up Robust Features (SURF) [12]. In [13], [14], the SURF points of interest also cover portions of the background, and their number and position are strongly influenced by the person's pose, thus making the final RE-ID less accurate. In our work, we used the same SURF descriptor but on points of interest at known positions to avoid that.

There are few studies in which descriptors for RE-ID are based only on 3D information [15]-[18]. In these works, the advantage of using descriptors based only on 3D allows their use even when changing clothes or under conditions of significant illumination change. However, this limits the final performance of RE-ID due to the limited and inaccurate information available. In [19], the depth features, such as normalized measurements of body parts, are calculated from the positions of joints. Their solution is limited by the joint extraction method not discussed in their article and, more generally, by the depth resolution.

To overcome these limitations, many works in the literature propose descriptors that integrate colour and 3D data to improve the RE-ID accuracy. In particular [20] combines clothing appearance descriptors with anthropometric measures extracted from depth data. Anthropometric measurements, such limb length, are strongly influenced by the method of joint extraction and the person's pose, so qualitative feedback on their estimation is insufficient to obtain a robust descriptor. Also, in [21], their descriptors do not consider the hand or ankle joints because, in contrast to our work, there is a lack of information about their uncertainty. Another example of the combined use of RGD-D data is [22]. The biometric data here involved are colour histograms split between the upper and lower body and on the subjects' height. In this case, the histograms consider the background reducing the final accuracy of RE-ID. In the case of occlusions, their method loses the height information, which is the only parameter they extract from the 3D information. More recent work [23] uses a single color-based descriptor generated from a partition grid, the size of which depends on skeletal posture obtained using 3D information. The points in each person's cloud are then grouped according to position in the grid. It requires an updated database for each posture, and the management of these coloured point clouds needs a higher computational cost than our approach, in which masks are used for a colour-based descriptor in each anatomical part from the position of the joints. In addition, because of the imperfect point cloud, some colour information may be lost, especially at the edges of human anatomical parts. Finally, using a single colour-based descriptor makes their RE-ID result less accurate.

All the previous works use direct strategies to extract features. The recent spread of deep learning has promoted the use of machine learning techniques to learn similarity models from training samples [24], [25]. However, learning-based strategies are time-consuming for training and testing steps and often unpredictable, which is why they are often used in industry only as support.

3. METHODOLOGY

The main steps of the proposed method are shown in Figure 1. Using a neural network, we first extrapolate the uncertainty ellipses related to the position of human joints. Then we select two descriptors, one based on colour, and one based on local feature, for pseudo-random points within the uncertainty ellipses of each joint. Afterwards, we compare the selected descriptors for each joint position to those stored in a spatiotemporal database used as ground truth. A sequential Monte Carlo method refines the joint positions for each descriptor to obtain the best matching score. Finally, the highest descriptor scores are concatenated to get the total RE-ID score.

We performed RE-ID only when a minimum number of 10 out of 18 joints is found and when the overlapping percentage of the bounding boxes of the selected person with the rest of the people is less than 30 %. In addition, since the RE-ID of people is also short-term, we assumed that people to be re-identified do not suddenly change clothes or accessories.

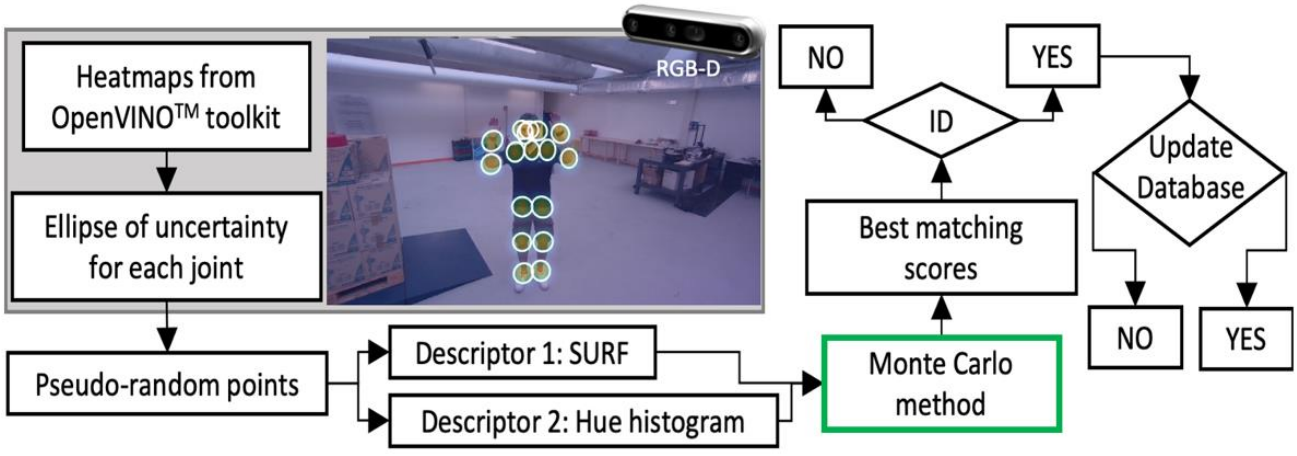


Figure 1. Flow chart of the proposed method.

3.1. Human pose estimation with joints uncertainty

To estimate the human pose, i.e., joints and connections between them for each person within an image, we applied a neural network provided by Intel in the OpenVINO toolkit, called *human-pose-estimation-0001*. OpenVINO toolkit is a collection of libraries for computer vision. It Enables CNN-based deep learning inference and contains an optimized version of OpenCV [26] libraries for Intel hardware. This network is based on OpenPose [27] approach with tuned MobileNet v1 [28] as a feature extractor. The network is also optimized to run on Inference Engine, which is a high-performance engine for neural networks developed by Intel; it allows inference on many Intel hardware such as Intel CPU, Processor Graphics, and FPGA.

The detection of humans carried out with this network is of the type Bottom-Up; this means that it does not detect all the human figures in the frame but looks for single interesting points and then, from them, tries to associate together parts/joints to obtain the pose associated to that human. With this net is possible to obtain up to 18 joints per person: ears, eyes, nose, neck, shoulders, elbows, wrists, hips, knees, and ankles. The network is validated on the COCO Dataset [29].

An RGB image is given as input to the network. The net produces two different outputs. The first consists of 18 probability maps, one for each joint, called heatmaps; the second consists of 19*2 layers, one for the horizontal and one for the vertical direction, called Part Affinity Fields (PAFs). With the first set of maps, it is possible to get all the joints in the image by obtaining the positions of the peaks. The second set, on the other hand, provides information on how to match the joints that correspond to a single person. In particular, by taking two possible points and the segment between them, it is possible to determine whether their pair belongs to the same person by checking whether the orientation and position of the segment on the frame match the orientation of the PAF unit vector. This is done by evaluating the scalar product of the segment orientation and PAF value at a discrete number of points within the same segment. If the orientation of the segment at these points is less than a threshold, the joint pair is saved. Finally, this network returns up to 18 joints for each person, assigning them the correct joints by listing pairs of valid joints that share a joint with another pair.

We used the heatmap of each joint to extrapolate the ellipse of uncertainty related to the position of the joints and modified the way of extracting the joints from the heatmaps. Associating an ellipse of uncertainty with their positions is even more important for how the OpenVINO neural network works for their extraction. It does not consider the joints' positions in previous frames, which causes their positions to change independently between frames.

To extract the ellipse of uncertainty of each joint, we first resize the probability map since the output map of the network is smaller than the input image. In Figure 2, there is an example of the non-symmetric probability distribution of a joint. Then we cut the probability map at a certain threshold, set it at 50 % probability, and obtain the contour area to select points for descriptors. The threshold percentage value affects the Cheesman constant [30], which changes the axes' dimensions of the ellipses and, accordingly, their final dimensions. Assuming a percentage higher than 50 % would result in a larger ellipses size, which for our purpose would result in a higher risk of considering the background instead of looking for the correct position of the joints. This percentage may change depending on the resolution of the image input to the neural network.

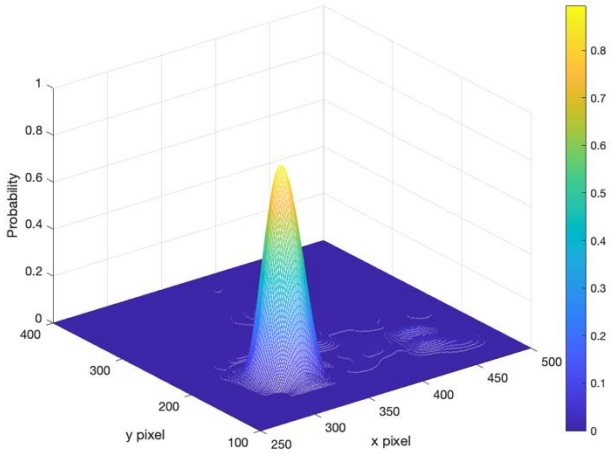
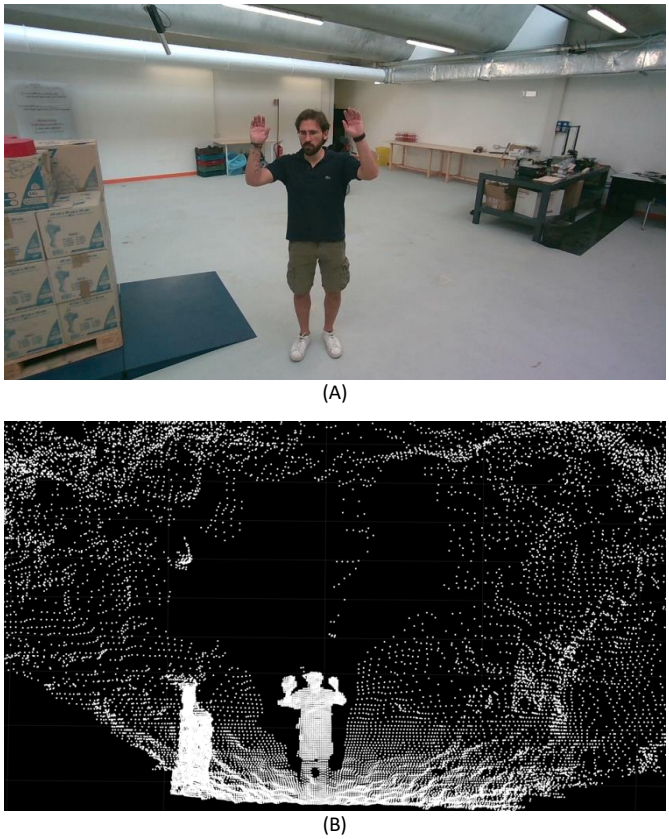


Figure 2. 3D example of non-symmetric probability distribution of the right wrist joint.

Before proceeding, we mixed the information from the 2D RGB image with the point cloud provided by the depth camera, Figure 3, allowing us to better estimate the position in the third dimension of each point within the selected area and remove those that are not on the surface of the human body through median filtering in the depth dimension. With the remaining points, we performed a weighted average with their weights given by the heatmap to find the new positions of each joint, i.e., their centres of mass, and around which to approximate a second-order surface. We discretely calculated the Hessian matrix by taking eight points around the new mean joint positions. Each term of the Hessian matrix can be approximated by taking the difference between two instances of the gradient vector



(A)

(B)

Figure 3. An example of an RGB (A) and depth image (B) (848 × 480 pixels) acquired by Realsense Depth Camera D455.

evaluated at nearby points of a generic joint i at row r_i and columns c_i.

The Hessian matrix terms are replaced by:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H_{11}} & \boldsymbol{H_{12}} \\ \boldsymbol{H_{21}} & \boldsymbol{H_{22}} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathbf{d^2}\boldsymbol{H}}{\mathbf{d}\boldsymbol{c^2}} & \dfrac{\mathbf{d}}{\mathbf{d}\boldsymbol{c}}\left(\dfrac{\mathbf{d}\boldsymbol{H}}{\mathbf{d}\boldsymbol{r}}\right) \\ \dfrac{\mathbf{d}}{\mathbf{d}\boldsymbol{r}}\left(\dfrac{\mathbf{d}\boldsymbol{H}}{\mathbf{d}\boldsymbol{c}}\right) & \dfrac{\mathbf{d^2}\boldsymbol{H}}{\mathbf{d}\boldsymbol{r^2}} \end{bmatrix}, \quad (1)$$

where

$$\frac{\mathrm{d}^2 H}{\mathrm{d}c^2} = \frac{\dfrac{H_{(r_i,c_i+n)} - H_{(r_i,c_i)}}{n} - \dfrac{H_{(r_i,c_i)} - H_{(r_i,c_i-n)}}{n}}{n}$$
$$= \frac{1}{n^2}\left(H_{(r_i,c_i+n)} - 2H_{(r_i,c_i)} + H_{(r_i,c_i-n)}\right)$$

$$\frac{\mathrm{d}^2 H}{\mathrm{d}r^2} = \frac{1}{n^2}\left(H_{(r_i+n,c_i)} - 2H_{(r_i,c_i)} + H_{(r_i-n,c_i)}\right)$$

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{\mathrm{d}H}{\mathrm{d}c}\right) = \frac{\mathrm{d}}{\mathrm{d}c}\left(\frac{\mathrm{d}H}{\mathrm{d}r}\right)$$
$$= \frac{1}{4n^2}\left(H_{(r_i+n,c_i+n)} - H_{(r_i+n,c_i-n)} - H_{(r_i-n,c_i+n)} + H_{(r_i-n,c_i-n)}\right)$$

with n the difference parameter in pixel. If n is too small there is a risk of considering high noise, while if n is too large, the result may be without meaning because it is averaged. For our application, we chose n = 4 pixels in accordance with the 9×9 box filters used in the SURF descriptor to approximate Gaussian second-order derivatives [12].

Finally, the uncertainty ellipse was evaluated with the covariance matrix to obtain a statistical description of the joint positions.

The covariance matrix corresponds to the negative of the inverse of the Hessian matrix [31]. Following this approach, we generate an object for each joint with 3 elements: pixel point, covariance matrix, and probability ellipse, Figure 4.

The results of the positions of the joints calculated through their centres of mass, as explained above, and through the peaks of the heatmaps, as output from the OpenVINO network, are shown in Figure 5. Figure 5 shows how, by extracting the position of each joint considering its statistical distribution from the heatmap and applying a depth filter to the selected points around it, the key points result more in the centre of human joints.

3.2. Descriptors matching

After the human pose estimation, finding the selected user in the frame is necessary. From the result of the joint detections, for each descriptor, we select a fixed number of pseudo-random points that follow a Gaussian probability distribution within the ellipse of uncertainty of each joint. To select these points, we first rotated each uncertainty ellipse using the respective eigenvector matrix. Then we translated them into the origin through the data of the positions of the centres of mass of the joints, so that the covariance matrices are symmetric and thus the eigenvectors are mutually orthogonal and the pseudo-random points to select could be considered independent along the two semi-axes of the ellipses. Once selected, the points were reprojected into the correct reference system.

After an initial analysis, we chose a colour-based histogram (HIST) descriptor and the SURF descriptor for our human signature. Each of the two descriptors compared with the ones stored in a database for comparison provides an independent result in terms of joint positions with the best match because to use the histogram-based descriptor, we approximated anatomical parts with geometric figures, which is a simplification not corresponding to reality.

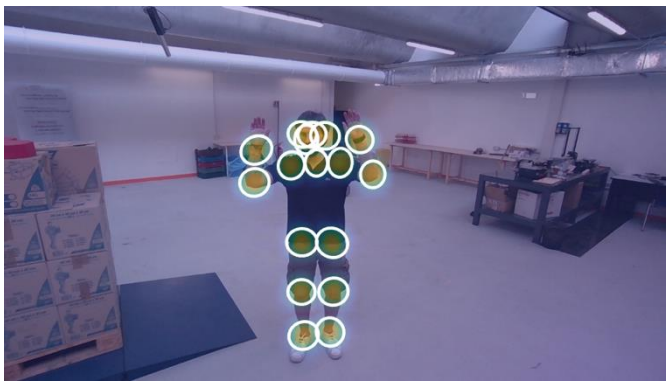
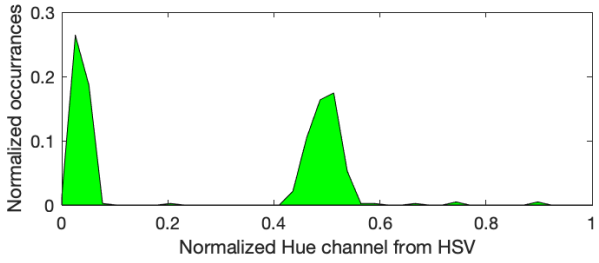


Figure 4. Heatmap results superimposed to the RGB image and the probability uncertainty ellipses estimated for each joint (50% confidence level with k=1.177 [30]).

In particular, the HIST descriptor is evaluated within the selected masks by analysing the Hue (H) channel of the HSV colour space. The mask was obtained for the face and torso by connecting joints on their edges. In contrast, for the upper and lower limbs, starting from the direction normal to each pair of joints, quadrilateral masks were found by adding additional points at a fixed length from the selected segment. This length is reduced or increased according to the user's distance from the camera. In the example of Figure 6, the width of the limb masks is set to 8 pixels. Before using the H-channel histogram as a descriptor, we analysed the most frequently chosen colour spaces, L*A*B* and HSV [32]. The H channel of the HSV colour space is the primary colour attribute and represents the phase angle of the colour, which is the usual name of the colour. Compared to L*A*B*, HSV encapsulates colour information in a way more similar to how the human eye perceives colour. Instead of defining a colour in terms of a combination of colour, the HSV colour model describes a colour with only one channel. In addition, H values in HSV are more robust to external light changes [33]. Also from our tests, where we compared the H-channel and the L*A* channels of HSV and L*A*B* colour spaces, respectively, the H-channel provides more robust behaviour in terms of light changes and colour description for processed images.

Figure 5. ROI from the initial RGB image with joint positions estimation using peak values (cross symbol in red) or centers of mass (plus symbol in green) of the heatmap.

(A)



(B)

Figure 6. Colour-based descriptor masks for each anatomical part superimposed on the Hue channel image (A); in green, an example of a selected anatomical part (A) with its histogram (B).

To express the similarity between the obtained histogram H_1 for each anatomical part and the one selected from the database H_2, we used the OpenCV correlation metric $S_H(H_1,H_2)$

$$S_H(H_1,H_2) = \frac{\sum_I (H_1(I)-\bar{H}_1)(H_2(I)-\bar{H}_2)}{\sqrt{\sum_I (H_1(I)-\bar{H}_1)^2 \sum_I (H_2(I)-\bar{H}_2)^2}}, \quad (2)$$

where

$$\bar{H}_k = \frac{1}{n_b}\sum_J H_k(J) \quad (3)$$

and n_b is the total number of histogram bins. For our test, we set n_b equal to 40 bins.

Only for the masks of the lower and upper limbs, we considered pseudo-random points around the selected joints because these limbs are the most prone to errors in joint placement, especially in the direction normal to the limb since there is a greater possibility of considering parts of the background as well. Each limb is divided into two anatomical regions for a total of eight anatomical parts. The pseudo-normal points for each joint at the end of each anatomical region are projected orthogonally only in the direction normal to the vector linking the two joints.

Instead, the SURF descriptor produces feature vectors of 64 floating elements for each selected point inside the uncertainty ellipse. The obtained descriptor vectors with similar values to those stored in the database are close in cosine similarity distance and far apart for different values. The cosine similarity S_C is defined as the cosine of the angle between two vectors, A and B, on dimension $\mathbb{R}^n$. The higher the cosine similarity, the higher the probability that the features are similar because the vectors are more aligned.

$$S_C(A,B) = \cos(\theta) = \frac{A\cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2}\sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

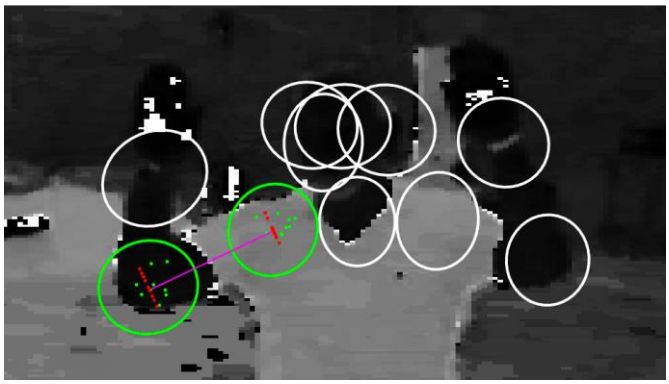
where A_i and B_i are components of vectors A and B respectively.

Matching for each descriptor is performed only for joints that exist for both the user in the frame being compared and the one in the database. With the constraints imposed for both descriptors, the working domain is 2D.

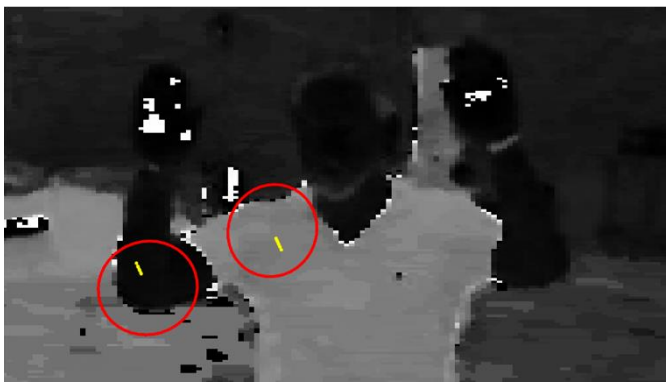
3.3. Monte Carlo Refinement

The selected pseudo-random points for each joint with their similarity scores ($S_H(H_1,H_2)$,$S_C(A,B)$) given by the matches with the database of the two descriptors are used as input for the Sequential Monte Carlo (SMC) method, also known as Particle Filter [34]. The idea behind SMC refinement for both descriptors is iteratively changing the joint positions inside the uncertainty ellipses to find the ones with a higher probability of matching in case the subject in the frame chosen to compare with the one in the database is the same.

The algorithms used for each of the two descriptors are reported as pseudocodes in Appendix A and explained below with an example. For the HIST and SURF descriptors, in this example, we considered the anatomical region of the right upper arm and left ankle, respectively. For the comparison, two identical frames were taken to be used both as ground truth for the database and as a frame to be compared. It allowed us to



(A)



(B)

Figure 7. An example of Monte Carlo refinement for the HIST descriptor of the right upper arm. In (A), with green dots, the pseudo-random points generated for each joint; red dots show their projection to the normal of the vector linking the two joints. Maximum HIST descriptor score equal to 0.88. The yellow dots in (B) show the new generated points after 3 iterations. New HIST descriptor score of 0.99.

ensure an exact match between the two frames with a maximum score of 1 for both descriptors.

For the HIST descriptor applied to the right upper arm are first generated 10 pseudo-random points for both the joint at the end of the anatomical area within the selected joint uncertainty ellipse and filtered in depth, Figure 7(A) green dots. All the pseudo-random points are then projected orthogonally in the direction normal to the vector linking the two joints, Figure 7(A), red dots. Moving to the left and right of each point by a fixed distance, in this case 8 pixels, the right upper arm is simplified into a rectangle and the histogram within the resulting mask is extracted. Instead, for the database is used the rectangle obtained from the initial positions of the joint mass centres and compared with each rectangle obtained from each pair of points.
Then, at each new step, SMC takes only half of the first pairs of points with the highest matching score with the database and generates new positions for these points. It moves each point left and right in the direction normal to the linking vector and with a distance proportional to the value of the ellipse's minor semi-axis with a trend of the logarithmic function in base 10: for each point, if the score found is low it moves far away from the previous position, instead if the score is high, it moves slightly away. If a newly generated point ends up outside the uncertainty ellipse, its position is not updated but remains the previous one. Each iteration saves the maximum score obtained from the database comparison until the maximum number of iterations is reached. Figure 7(B) shows the motion of the joints' position after 3 iterations. The maximum match for the HIST descriptor passes from 0.88 to 0.99.

A similar approach is used for the SURF descriptor. In this case, we did not consider two joints at a time but a single joint with possible movements in both directions. Figure 8 shows an example of 20 pseudo-random points generated for the left ankle within the selected joint uncertainty ellipse. The SURF descriptor scores were obtained by comparing each point with the descriptor associated with the joint centre of mass used for the database. Then, at each new step, new positions are generated starting from the previous ones, moving them in a random direction and a distance proportional to the value of the ellipse's minor semi-axis with a logarithmic trend. As seen in Figure 8(B), after 7 iterations, the points moved to the areas of the highest match from an initial maximum match of 0.94 to 0.99. Around the maximum peak, the matching decreases while still having local disturbance peaks, making it impossible to predict the surface pattern with a polynomial.

For both descriptors, such a high score already at the first step is also due to the fact that the same frame was used in this example for the database and the one to compare. For the same reason, the score obtained with SMC refinement is very close to 1.

Once the highest match is found for each anatomical part and each joint for HIST and SURF descriptors, the final score is obtained by concatenating these results for each person with the following expressions:

$$\boldsymbol{S_{\mathrm{HIST}}} = \frac{||\,\vec{\boldsymbol{S}}_{\boldsymbol{H}}(\boldsymbol{H_1}, \boldsymbol{H_2})||}{\sqrt{\boldsymbol{n_k}}}\ ;\ \boldsymbol{S_{\mathrm{SURF}}} = \frac{||\,\vec{\boldsymbol{S}}_{\boldsymbol{C}}(\boldsymbol{A}, \boldsymbol{B})||}{\sqrt{\boldsymbol{m_k}}} \tag{5}$$

where $\vec{S}_H(H_1, H_2)$ is the vector containing all pairs of points with the highest matches obtained with the HIST descriptor, $\vec{S}_C(A, B)$, on the other hand, contains a point for each joint with the highest matches obtained with the SURF descriptor; n_k and m_k are the number of anatomical parts and joints found with the HIST and SURF descriptor, respectively. Finally, the total similarity score obtained for person ID is the average of the two scores S_{HIST} and S_{SURF}.

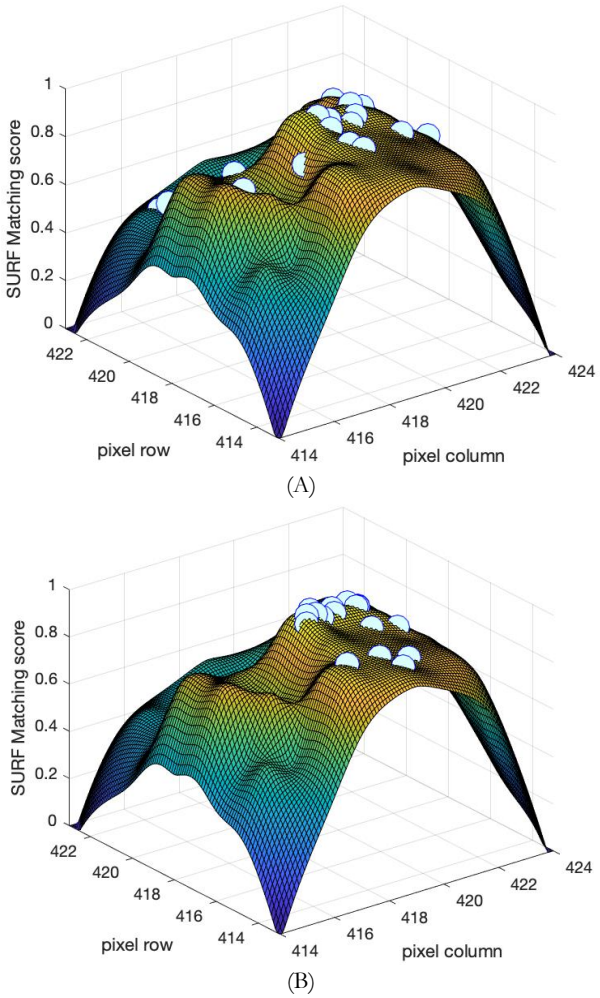


Figure 8. The mesh is an example of the SURF matching score values for all points around the selected joint with the one of the database; the 20 pseudo-normal points with their SURF Matching values at initial configuration (A) with a maximum score of 0.94 and after 7 iterations (B) of sequential Monte Carlo method with a maximum score of 0.99.

When a user is identified, the descriptor vectors to be used as ground truth for the database are updated; or when the angle between limbs or light variations exceed 30 %. We monitor the light with the L* channel of L*A*B* spatial color. In addition, descriptor vectors for the database are calculated using the position of the centers of mass of each joint. All databases are also classified based on 3D information about user orientation. The database size selected for each user orientation in our application was set to 10 vectors for both descriptors.

4. VALIDATION

We assessed the performance of the proposed RE-ID method by first evaluating the effects of Monte Carlo refinement and then the RE-ID results obtained with the whole method. The whole validation was implemented in MATLAB 2020b.

We carried out the assessment on the most widely used dataset for re-identifying people with RGB-D images, such as the BIWI RGBD-ID dataset [35]. Although it is an RGB-D dataset of people targeted to long-term people RE-ID, we selected 28 people in the training sets in which, for each person, about 300 images were collected on the same day and in the same scene, so

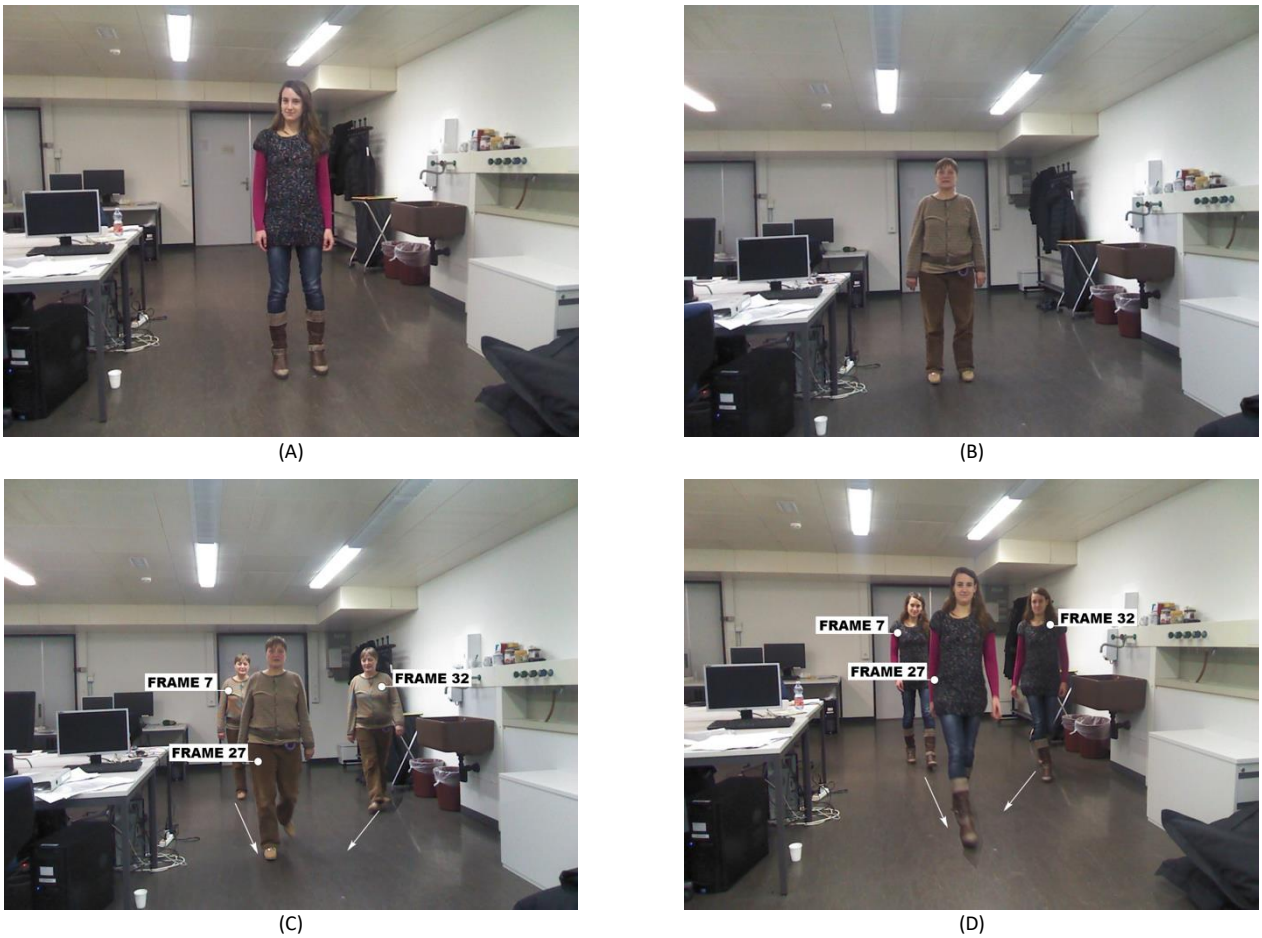
that the people were dressed in the same way. People moved slightly in the "Still set", while in the "Walking set", each person walked from different viewing angles. In addition, the same number and subject ID are present in both sets.

4.1. Monte Carlo Refinement

To show the performance of the SMC refinement, we tested its effects on both descriptors. We showed an example of the results of the scores obtained with each descriptor using the single image in Figure 9(A) from a "Still set" of the BIWI RGBD-ID dataset as database. As frames for comparison, however, we used the remaining frames from the "Still set" of the user in Figure 9(A), another "Still set" from a different user in Figure 9(B), and their "Walking set", Figure 9(C) and Figure 9(D), respectively of the two users. We also tested the effect of SMC refinement by comparing the database user with another user's dataset to verify that, in the wrong case, the SMC refinement does not increase the final scores in a way that would compromise RE-ID.

Figure 10 shows the results of the HIST descriptor. The results are obtained by concatenating the scores from each limb but not considering the torso and face masks to consider only the anatomical areas where SMC refinement acts. By concatenating the final score also with the missing areas, mainly with the torso area, the final solution can only be better and more robust because of the size and location of that area. In particular, Figure 10(A) shows the results obtained by comparing the database image of Figure 9(A) with the "Still set" in which the user is the same or the wrong one. As shown in Figure 10(A), in the case of the correct user, the SMC refinement improved the final descriptor score in each frame. However, it is high even without the refinement because the user moved slightly in this dataset, and the lighting conditions can be considered constant. In Figure 10(B), the same database image is compared with the "Walking set". In this case, the contribution of refinement is greater because of the user's movement in front of the camera and from a different angle. It demonstrates how refinement is possible with this descriptor, even if lighting conditions and user kinematics change and even with blurred frames. In particular, frame 32 in Figure 9 (C) and Figure 9 (D) corresponds to the moment when users start walking from a different viewing angle. In both cases in Figure 10, comparing the user with the wrong one, refinement increases the final descriptor score but in a way that does not compromise the RE-ID. It is also true that the lower the score, the easier it is to find a better solution by performing SMC refinement. The importance is to keep this increment controlled, as in our case.

In contrast, Figure 11 shows the results obtained with the SURF descriptor by concatenating the results of each joint. This descriptor is strongly influenced by user kinematics. Figure 11(A) and Figure 11(B) refer to a "Still set" with the same user of the database and one with the wrong user, respectively. Because of the slight movements of the user, comparing all frames with the single image selected for the database of Figure 9(A) is acceptable. If, on the other hand, we compare the database image with a "Walking set", Figure 11(C), even if the user is the same between the database and dataset, the scoring trend after the



(A) (B) (C) (D)

Figure 9. (A) frame selected from "Still set" to be used as database for comparison; (B) is an example of a frame from "Still set" of another user with respect to the one saved in the database; (C)-(D) examples of frames from "Walking set" of both users.

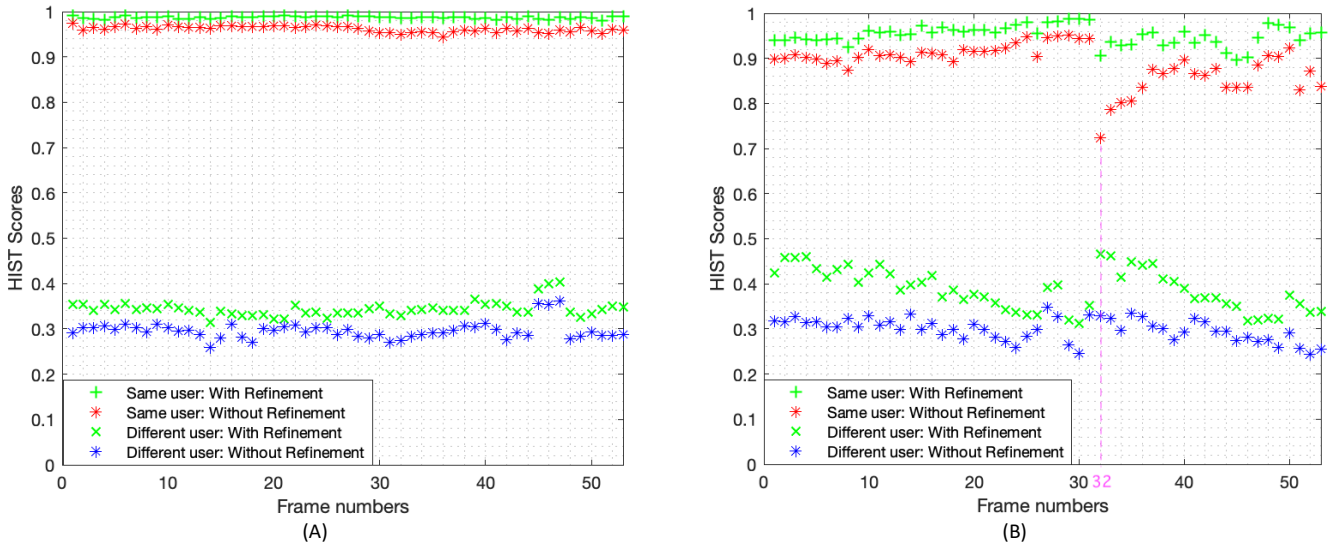


Figure 10. Results of the HIST descriptor score for each frame using the "still set" (A) and "Walking set" (B), in both cases comparing the same users (red asterisks) or the wrong one (blue asterisks) without the SMC refinement or with it (green symbols).

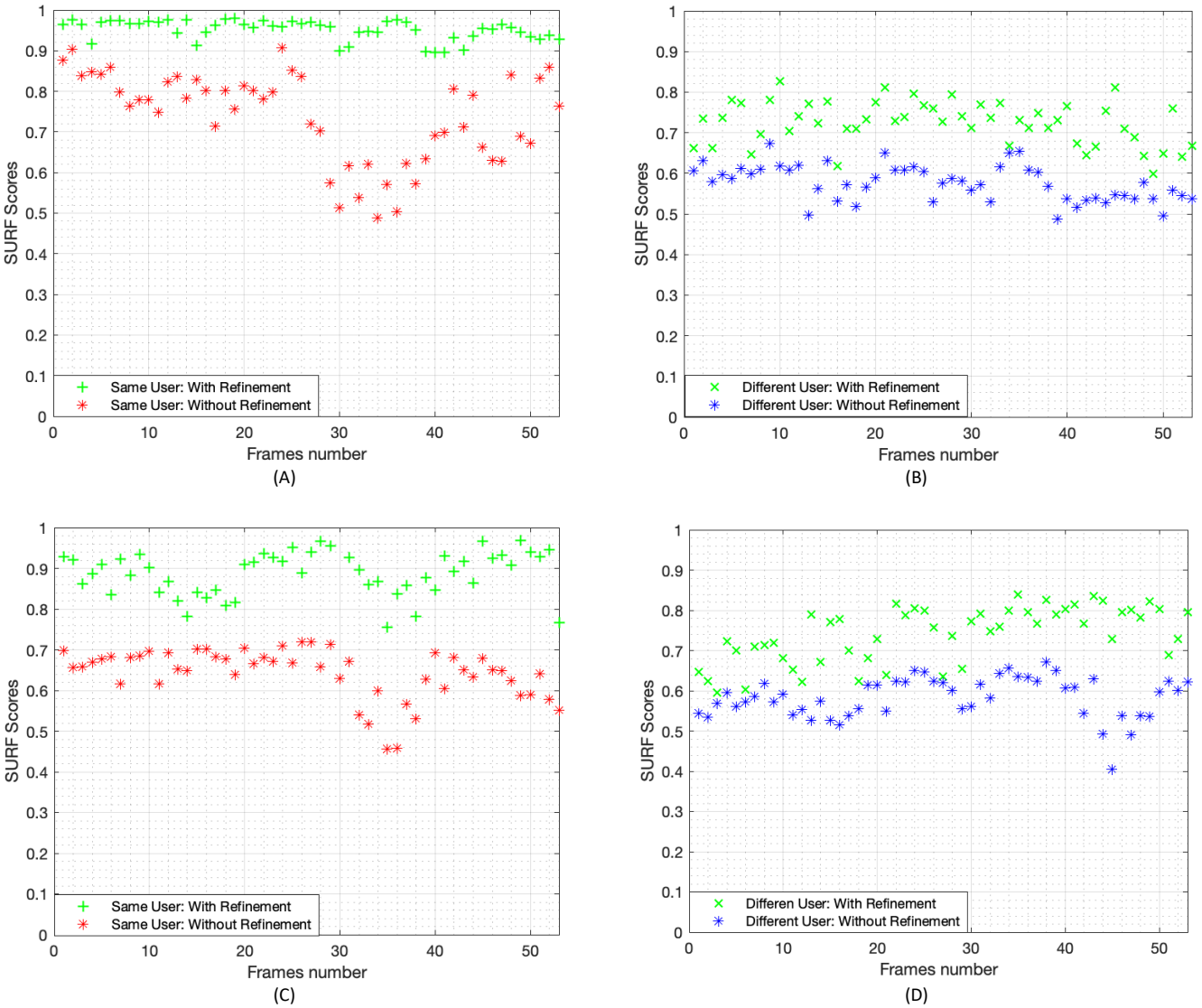


Figure 11. Results of the SURF descriptor score for each frame using the "still set" (A)-(B) and "Walking set" (C)-(D). (A)-(C) are obtained by comparing the same users (red asterisks) and (B)-(D) with the wrong one (blue asterisks). All the scores in green are obtained with SMC refinement.

refinement is up and down because comparing the same database image for all possible kinematic poses of the user is a wrong assumption. In this case, the only acceptable comparison can be made for the first few frames where the user is farther from thecamera, but his kinematics does not change. To overcome this problem, we use multiple database images stored in the proposed method, considering not only the user's orientation but also the user's kinematics.

4.2. RE-ID ranks

We assessed the performance of the proposed RE-ID method using the Cumulative Matching Characteristics (CMC) metric [36], which is commonly used for evaluating RE-ID algorithms. For every k from 1 to the number of training subjects, the CMC expresses the average person recognition rate computed when the correct person is included in the k best classification scores (rank-k). The most relevant ranks for RE-ID are the first ranks. In particular, the rank-1 matching rate refers to the probability that the probe image and the image that ranks first in similarity in the search gallery belong to the same subject. Therefore, a higher recognition rate at rank-1 means the correct ID of the subject in the highest rank and thus better model performance since the chance of having false positives with incorrect subject ID is low.

We selected the "Still set" of images from BIWI RGBD-ID as the probe, while the "Walking set" as the gallery (database). The matching for every subject in the probe with respect to the gallery provides a ranking, Figure 12. This procedure is repeated for every subject and then averaged to obtain the final recognition rate.

On the same dataset, we ran the RE-ID OpenVINO demo [37] to compare with our method in terms of CMC curves. The OpenVINO demo detects pedestrians in the frames through the pedestrian detector network *person-detection-retail-0013* and re-identifies them through the pedestrian RE-ID model for a general scenario *person-reidentification-retail-0277*. The choice of this pre-trained model for the RE-ID is backed by the superior Market-1501 rank-1 [38] accuracy of 96.2 %. This demo was modified to use the RE-ID network to compare each probe set frame with all the gallery frames.

The CMC curves obtained by our method and the RE-ID of OpenVINO are shown in Figure 13.

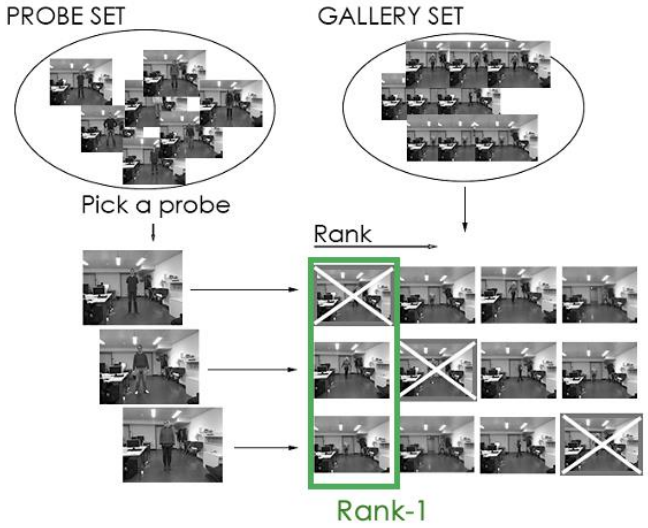


Figure 12. Cumulative Matching Characteristic metric used as validation.

Both curves in Figure 13 showed a high recognition rate at rank-1. This is due to the goodness of the methods but also to the way the images of the BIWI RGBD-ID dataset were acquired: the camera was placed in the same position and fixed throughout the time of the tests; also, the same background was used between the probe and gallery images. However, our method gave a better result than that obtained with OpenVINO: 99.1 % vs 98.4 %. This difference can be increased by changing the background between the probe set and the gallery set since the descriptor extracted by the OpenVINO neural network for RE-ID uses a whole-body image as input and thus goes to consider parts of the background as well.

In addition, CMC over-penalizes false positives while ignoring missed RE-IDs. The decision of whether the selected subject is present in the gallery depends on the choice of the similarity threshold used. In our method, this threshold can be set to a high value without missing RE-IDs, thanks to the SMC refinement process of matching descriptors.

5. CONCLUSIONS

In this paper, an accurate method for the RE-ID of people was presented. Using RGB-D data as a combination of soft biometric traits for similarity matching, we extrapolated a color-based descriptor and a local feature descriptor considering the uncertainty of human joints.

In contrast to the OpenVINO RE-ID approach, our method does not require any effort for the learning steps, and therefore the features extracted do not depend on the set of images on which the training was performed. It is also more stable than a learning-based strategy because it uses direct strategies to extract features.

Thanks to its properties, its use is possible in environments where incorrect ID is not allowed because it would increase the risk to operators, such as in industrial environments. With our method, this constraint keeps the RE-ID rate high because, due to the high similarity value obtained by sequential Monte Carlo refinements in descriptor matching, we can use high similarity thresholds without risking not identifying the selected subject.

We achieved a 99.1 % recognition rate at rank-1 on the BIWI RGBD-ID public dataset. This result represents a significant improvement over previous approaches in the literature obtained using the same public dataset.

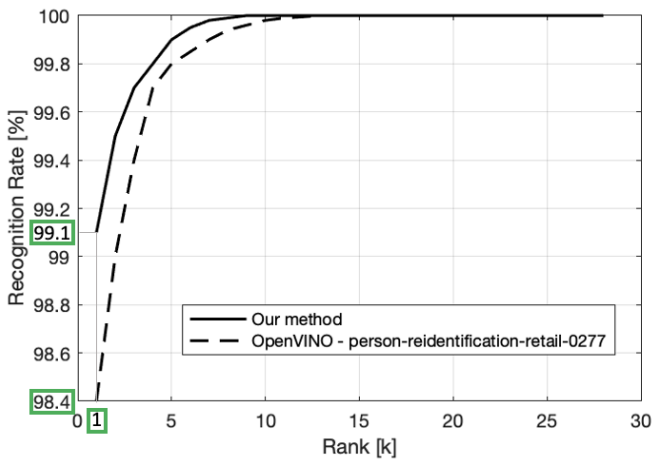


Figure 13. Cumulative Matching Characteristic curves obtained by our re-identification method and OpenVINO neural network on BIWI RGBD-ID dataset.

REFERENCES

[1] S. U. Khan, T. Hussain, A. Ullah, Sung Wook Baik, Deep-ReID: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance, Multimedia Tools and Applications Springer, 2021, pp. 1-22.
DOI: 10.1007/s11042-020-10145-8

[2] E. Yaghoubi, K. Aruna, P. Hugo, Sss-pr: A short survey of surveys in person re-identification. Pattern Recognition Letters 143, 2021, pp. 50-57.
DOI: 10.1016/j.patrec.2020.12.017

[3] B. Wang, G. Wang, K. L. Chan, L. Wang, Tracklet association with online target-specific metric learning, Proc. of the IEEE Conf. on computer vision and pattern recognition, Columbus, OH, USA , 23-28 June 2014, pp. 1234-1241.
DOI: 10.1109/CVPR.2014.161

[4] N. An, S. Y. Sun, X. G. Zhao, Z. G. Hou, Online context-based person re-identification and biometric-based action recognition for service robots, IEEE 29th Chinese Control and Decision Conference (CCDC), Chongqing, China, 28-30 May 2017, pp. 3369-3374.
DOI: 10.1109/CCDC.2017.7979088

[5] A. Luchetti, A. Carollo, L. Santoro, M. Nardello, D. Brunelli, P. Bosetti, M. De Cecco, Human identification and tracking using ultra-wideband-vision data fusion in unstructured environments, Acta IMEKO, 10(4), 2021, pp. 124-131.
DOI: 10.21014/acta_imeko.v10i4.1139

[6] K. Koide and J. Miura, Identification of a specific person using color, height, and gait features for a person following robot, Robotics and Autonomous Systems, 2016, pp. 76-87.

[7] Intel Corp., OpenVINO toolkit. Online [Accessed 28 Nov 2022] https://docs.openvinotoolkit.org/latest/index.html

[8] A. Dantcheva, C. Velardo, A. D'angelo, J.L. Dugelay, Bag of soft biometrics for person identification, Multimedia Tools and Applications, 51(2), 2011, pp. 739-777.
DOI: 10.1007/s11042-010-0635-7

[9] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Transactions on pattern analysis and machine intelligence, 35(7), 2012, pp. 1622-1634.
DOI: 10.1109/TPAMI.2012.246

[10] R. F. de Carvalho Prates, W. R. Schwartz, CBRA: Color-based ranking aggregation for person re-identification, IEEE Int. Conf. on Image Processing (ICIP), Quebec City, QC, Canada, 27-30 September 2015, pp. 1975-1979.
DOI: 10.1109/ICIP.2015.7351146

[11] X. Liu, H. Wang, Y. Wu, J. Yang, M. H. Yang, An ensemble color model for human re-identification, IEEE Winter Conf. on Applications of Computer Vision, Waikoloa, HI, USA, 05-09 January 2015, pp. 868-875.
DOI: 10.1109/WACV.2015.120

[12] H. Bay, T. Tuytelaars, L. V. Gool, Surf: Speeded up robust features, European Conf. on computer vision, Springer, 2006, pp. 404-417.
DOI: 10.1007/11744023_32

[13] O. Hamdoun, F. Moutarde, B. Stanciulescu, B. Steux, Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, Second ACM/IEEE Int. Conf. on Distributed Smart Cameras, Palo Alto, CA, USA, 07-11 September 2008, pp. 1-6.
DOI: 10.1109/ICDSC.2008.4635689

[14] M. I. Khedher, M. A. El-Yacoubi, B. Dorizzi, Probabilistic matching pair selection for surf-based person re-identification, Proc. of the Int. Conf. of Biometrics Special Interest Group (BIOSIG), ISSN: 1617-5468, Darmstadt, Germany, 06-07 September 2012, pp. 1-6.

[15] S. Gharghabi, F. Shamshirdar, T. A. Shangari, F. Maroofkhani, People re-identification using 3D descriptor with skeleton information, Int. Conf. on Informatics, Electronics & Vision (ICIEV), Fukuoka, Japan, 15-18 June 2015, pp. 1-5.
DOI: 10.1109/ICIEV.2015.7333986

[16] A. Wu, W. S. Zheng, J. H. Lai, Robust depth-based person re-identification, IEEE Transactions on Image Processing, 26(6), 2017.
DOI: 10.48550/arXiv.1703.09474

[17] M. Munaro, A. Basso, A. Fossati, L. Van Gool, E. Menegatti, 3D reconstruction of freely moving persons for re-identification with a depth sensor, IEEE Int. Conf. on robotics and automation (ICRA), Hong Kong, China, 2014, pp. 4512-4519.
DOI: 10.1109/ICRA.2014.6907518

[18] S. Cosar, C. Coppola, N. Bellotto, Volume-based Human Re-identification with RGB-D Cameras, VISIGRAPP (4: VISAPP), 2017, pp. 389-397.
DOI: 10.5220/0006155403890397

[19] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, V. Murino, Re-identification with rgb-d sensors, European Conf. on Computer Vision, Springer, Berlin, Heidelberg, 2012, pp. 433-442.
DOI: 10.1007/978-3-642-33863-2_43

[20] F. Pala, R. Satta, G. Fumera, and F. Roli, Multimodal person reidentification using rgb-d cameras, IEEE Transactions on Circuits and Systems for Video Technology, 26(4), 2016, pp. 788–799.
DOI: 10.1109/TCSVT.2015.2424056

[21] M. Munaro, S. Ghidoni, D. T. Dizmen, E. Menegatti, A feature-based approach to people re-identification using skeleton keypoints, IEEE Int. Conf. on robotics and automation (ICRA), Hong Kong, China, 2014, pp. 5644-5651.
DOI: 10.1109/ICRA.2014.6907689

[22] A. Møgelmose, T.B. Moeslund, K. Nasrollahi, Multimodal person re-identification using RGB-D sensors and a transient identification database, Int. Workshop on Biometrics and Forensics (IWBF), Lisbon, Portugal, 04-05 April 2013, pp. 1-4.
DOI: 10.1109/IWBF.2013.6547322

[23] C. Patruno, R. Marani, G. Cicirelli, E. Stella, T. D'Orazio, People re-identification using skeleton standard posture and color descriptors from RGB-D data. Pattern Recognition, 89, 2019, pp. 77-90.
DOI: 10.1016/j.patcog.2019.01.003

[24] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE transactions on pattern analysis and machine intelligence, 44(6), 2021.
DOI: 10.1109/TPAMI.2021.3054775

[25] D. Wu, S. J. Zheng, X. P. Zhang, C. A. Yuan, F. Cheng, Y. Zhao, Y-J Lin, Z. Q. Zhao, Y. L. Jiang, D. S. Huang. Deep learning-based methods for person re-identification: A comprehensive review. Neurocomputing, 337, 2019, pp. 354-371.
DOI: 10.1016/j.neucom.2019.01.079

[26] G. Bradski, The openCV library. Dr. Dobb's Journal: Software Tools for the Professional Programmer, 25(11), 2000, pp. 120-123.

[27] Z. Cao, T. Simon, S. E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, Proc. of the IEEE Conf. on computer vision and pattern recognition, Honolulu, 2017, pp. 7291-7299.
DOI: 10.1109/CVPR.2017.143

[28] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.
DOI: 10.48550/arXiv.1704.04861

[29] COCO Consortium, Common Objects in Context. Online [Accessed 28 Nov 2022]
https://cocodataset.org/#home

[30] R. C. Smith, P. Cheeseman, On the representation and estimation of spatial uncertainty, The international journal of Robotics Research, 5(4), 1986, pp. 56-68.
DOI: 10.1177/027836498600500404

[31] W. C. Thacker, The role of the Hessian matrix in fitting models to measurements, Journal of Geophysical Research: Oceans, 94(C5),

1989, pp. 6177-6196.
DOI: 10.1029/JC094IC05P06177

[32] M. W. Schwarz, W. B. Cowan, J. C. Beatty, An experimental comparison of RGB, YIQ, LAB, HSV, and opponent color models, Acm Transactions on Graphics (tog), 6(2), 1987, pp. 123-158.
DOI: 10.1145/31336.31338

[33] P.A. Marin-Reyes, J. Lorenzo-Navarro, M. Castrillón-Santana, Comparative study of histogram distance measures for re-identification, arXiv preprint arXiv:1611.08134, 2016.
DOI: 10.48550/arXiv.1611.08134

[34] A. Doucet, N. De Freitas, J. G. Neil Gordon, Sequential Monte Carlo methods in practice. Ed. Arnaud Doucet. Vol. 1. No. 2. New York: springer, 2001.
DOI: 10.1007/978-1-4757-3437-9

[35] M. Munaro, A. Fossati, A. Basso, F. Menegatti, L. Van Gool, One-shot person re-identification with a consumer depth camera, In Person Re-Identification, Springer, London, 2014, pp. 161-181.
DOI: 10.1007/978-1-4471-6296-4_8

[36] H. Moon, P. J. Phillips, Computational and performance aspects of PCA-based face-recognition algorithms, Perception, 30(3), 2001, pp. 303-321.
DOI: 10.1068/p2896

[37] Intel Corp., Pedestrian tracker c++ demo. Online [Accessed 28 Nov 2022] https://docs.openvino.ai/latest/omz_demos_pedestrian_tracker_demo_cpp.html

[38] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, Proc. of the IEEE Int. Conf. on computer vision, Santiago, Chile, 07-13 December 2015, pp. 1116-1124.
DOI: 10.1109/ICCV.2015.133

APPENDIX A

Sequential Monte Carlo: SURF

Input: P_{CM} (Joint center of mass)
$P_i \dots P_N$ (Initial pseudo-random points)
$S_i(P_i) \dots S_N(P_N)$ (Descriptor matching scores)
Output: New point $(\hat{P}_i)$ with S_{MAX}

for $k \leftarrow 1$ to Max number of iterations **do**
  for $i \leftarrow 1$ to N **do**
    while $\hat{P}_i$ outside ellipse of uncertainty **do**
      DIST $\leftarrow |Ellipseminorsemiaxis * \log(S_i)|$
      If DIST $> Ellipseminorsemiaxis$ **then**
        DIST $= Ellipseminorsemiaxis$
      end if
      ROT $\leftarrow$ random$[0°, 360°]$
      $\hat{P}_i \leftarrow P_i$ + Distance ($DIST$) in Direction (ROT)
    end while
  Calculate descriptor matching score: $S_i(\hat{P}_i)$
  end for
$\vec{S}_k = max(S_i(\hat{P}_i) \dots S_N(\hat{P}_N))$
end for
return $S_{MAX} = max\ (\vec{S})$

Sequential Monte Carlo: HIST

Input: P^I_{CM}, P^{II}_{CM} (Center of mass of I and II joint)
$P^I_i \dots P^I_N$ (Initial pseudo-random points I joint)
$P^{II}_j \dots P^{II}_N$ (Initial pseudo-random points II joint)
$S_i(P^I_i, P^{II}_i) \dots S_N(P^I_j, P^{II}_j)$ (Descriptor matching scores)
Output: New points $(\hat{P}^I_i, \hat{P}^{II}_j)$ with S_{MAX}

ROT $\leftarrow$ normal direction to $\overline{P^I_{CM}\ P^{II}_{CM}}$
$P^I_i \dots P^I_N = P^I_i \dots P^I_N * ROT$
$P^{II}_j \dots P^{II}_N = P^{II}_j \dots P^{II}_N * ROT$

for $k \leftarrow 1$ to Max number of iterations **do**
  for $i \leftarrow 1$ to N **do**
    for $j \leftarrow 1$ to N **do**
      Saved $S_{ij}(P^I_i, P^{II}_j)$ in matrix $S_{TOT}[i,j]$
    end for
  end for
  $sort(S_{TOT}) \rightarrow S_{TOT}[1..\frac{N}{2}]$
  DIST(i, j) $\leftarrow |Ellipseminorsemiaxis * \log(S_{ij})|$
  If DIST $> Ellipseminorsemiaxis$ **then**
    DIST $= Ellipseminorsemiaxis$
  end
  for $i, j \leftarrow 1$ to N **do**
    $\hat{P}^I_i \leftarrow P^I_i \pm$ Distance ($DIST$) in Direction (ROT)
    $\hat{P}^{II}_j \leftarrow P^{II}_j \pm$ Distance ($DIST$) in Direction (ROT)
    Calculate descriptor matching score: $S_i(\hat{P}^I_i, \hat{P}^{II}_j)$
  end for
  $\vec{S}_k = max(S_{ij}(\hat{P}^I_i, \hat{P}^{II}_j) \dots S_{NN}(\hat{P}^I_N, \hat{P}^{II}_N))$
end for
return $S_{MAX} = max\ (\vec{S})$