IJCV manuscript No.
(will be inserted by the editor)

Mapping Appearance Descriptors on 3D Body Models for People Re-identification

Davide Baltieri · Roberto Vezzani · Rita Cucchiara

Received: date / Accepted: date

Abstract People Re-identification aims at associating multiple instances of a person's appearance acquired from different points of view, different cameras, or after a spatial or a limited temporal gap to the same identifier. The basic hypothesis is that the person's appearance is mostly constant. Many appearance descriptors have been adopted in the past, but they are often subject to severe perspective and view-point issues. In this paper, we propose a complete re-identification framework which exploits non-articulated 3D body models to spatially map appearance descriptors (color and gradient histograms) into the vertices of a regularly sampled 3D body surface. The matching and the shot integration steps are directly handled in the 3D body model, reducing the effects of occlusions, partial views or pose changes, which normally afflict 2D descriptors. A fast and effective model to image alignment is also proposed. It allows operation on common surveillance cameras or image collections. A comprehensive experimental evaluation is presented using the benchmark suite 3DPeS.

Keywords 3D human model, People Re-identification

1 Introduction

People surveillance is an important topic in computer vision and pattern recognition research. Surveillance tasks should be as fast as possible in knowledge extraction to allow rapid crime prevention or first aid intervention. Similarly, in multimedia forensics, video archives must be mined as quickly as possible. Detecting whether a person selected as query has already been observed in a different place and at a different time in the camera network or in a video archive remains a challenging task in surveillance and forensics application. In the computer vision community, this problem is well known and called *People Re-Identification*. People Re-identification is also a fundamental task for the analysis of long-term activities and behaviors of specific people and for the latest studies on people social interaction.

The importance of this field has been clearly pointed out in the recent survey on re-identification methods [1], in the book on "Person Re-Identification" edited by Gong *et al.* [2], and has been confirmed by the huge amount of papers published on the topic. Although it is subject to a basic hypothesis of real appearance constancy in the observed time interval – i.e., no camouflage are allowed – the acquired appearance can vary considerably because of people orientation, illumination conditions, presence of occlusions and so on. In addition, due to a potentially high number of similar people (e.g., people wearing similar clothes) often no accurate temporal or spatial constraints can be exploited.

In this paper, we present the design of a complete system for people re-identification based on the mapping of appearance descriptors to 3D body models (called SARC3D). The adoption of 3D body models is quite new for re-identification, as opposed to other computer vision fields, such as for example motion capture and posture estimation [3, 4]. The challenges connected with 3D models rely on the need to obtain accurate people detection, segmentation and estimation of the 3D orientation for correct model-to-image alignment.

The key points of this work are:

Davide Baltieri, Roberto Vezzani, Rita Cucchiara
Dipartimento di Ingegneria Enzo Ferrari- University of Modena and Reggio Emilia, Via Vignolese, 905 - 41125 Modena - Italy
E-mail: davide.baltieri@gmail.com, roberto.vezzani@unimore.it, rita.cucchiara@unimore.it

1. the definition of a non-articulated 3D body model for people re-identification;
2. the selection of color and texture descriptors to generate a model-based person signature;
3. the definition of a suitable model distance;
4. the description of the overall framework, including model-to-image alignment and mapping.

The rest of the paper is organized as follows. Section 2 describes the state of the art and, in particular, the drawbacks of using traditional signatures. The importance and the advantages of using 3D body models are described in Section 3, together with a an outline of the proposed framework. Section 4 provides a mathematical definition of the SARC3D model and the related algorithms proposed to generate the person signature. The multi-shot integration step is detailed in Section 5, where two different approaches are provided: one is aimed at reaching the best performance and the other one at reducing the descriptor size. Section 6 contains the formulation of a suitable distance metric, required to compare and match persons' signatures. Qualitative and quantitative results of the framework as well as some comparisons with the state of the art approaches are reported in Section 7. Finally, some concluding remarks and future research directions are given in Section 8.

2 Related Works

People re-identification is a matching problem when it comes to items having the same, or at least similar, shape and structure. As a consequence, most of the techniques proposed are based on appearance features such as color and texture rather than shape and geometrical properties, with all the pros and cons involved with the discriminating capabilities of these features. For exhaustive surveys on appearance methods for re-identification, please refer to the works by Doretto *et al.* [5], Vezzani *et al.* [1], and Gong *et al.* [2].

The first approaches proposed were based on **global descriptors**, since feature de-localization usually means view invariance. Among others, global color histograms [6,7] or texture features [8,9] have been mainly implemented, which commonly resulted in confusing people wearing clothes with similar colors.

To cope with this limitation, generic or custom body models have been introduced, allowing us to associate each appearance descriptor with a specific spatial location. These models do not need to be articulated and very precise as in motion capture or action analysis scenarios; however, a model-based localization provides more coherent and representative descriptions and allows a correct comparison of corresponding body parts. Problems due to occlusions and different view points are thus minimized.

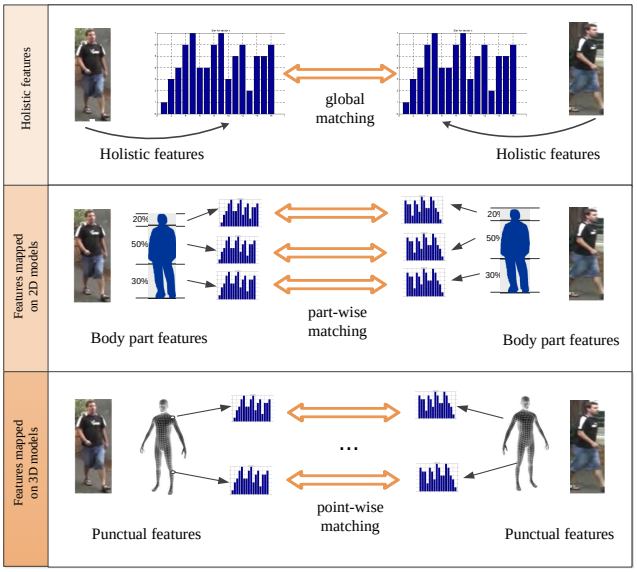


Fig. 1 Three different re-identification approaches. First row: global features are computed on the whole person detected. Second row: 2D body models allow generation of a specific signature for each body part, which is usually an horizontal slice. Third row: 3D body models allow generation of punctual features, mapped to view-independent body locations.

Simplified **2D models** are the most commonly used (see Fig. 1). Among others, in surveillance and forensics field the most widely used hypothesis for a body model is the *cylindrical* shape and the *legs-torso-head* structure. The horizontal variations of appearance are neglected by modeling a person as a cylindrical shape (or more generally as a solid of revolution). Color or texture distributions along the vertical axis are the only important data. For example, in [10] the person mask is divided into ten horizontal stripes and the mean color of each stripe is stored as a representative feature.

In the legs-torso-head model, instead, the target silhouette is divided into three horizontal parts, ideally corresponding to the legs (and thus to the pants/skirt appearance), torso (i.e., shirt or jacket) and head (i.e., hair). This segmentation can be obtained using fixed sizes [8,11]: for example, [8] places the cuts at 30% and 80% of the total height, [11] at 15% and 70%.

Other methods do not divide the detected region into fixed parts but propose adaptive solutions. For example, the authors of [12] automatically compute the cut points through profile histograms. Moreover, they further split the torso and the legs into two parts using a symmetry-based algorithm. Finally, [13] adopts a body part detector to extract the position of each limb,

torso and head. A high quality data source is required in this case and the body part detector is computationally expensive.

The most promising solutions based on 2D human models have been proposed in the last years by [12] and [5]. The first approach [12] uses human vertical symmetry to partition appearance images into five regions. For each region, three features describing complementary aspects of human appearance are extracted: the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent textures; a nearest neighbor matching schema is then applied. The work by [5], instead, reviews several appearance descriptors and then proposes a part-based signature, which outperforms global descriptors in the tests provided.

Rather than working on new features and signature types, most proposals have been recently focused on distance metric learning. Machine learning approaches have been adopted to project feature vectors into a space where intra-person distances are reduced while inter-person ones are increased. In other words, the distance learning step should select the best subset or the best combination of computed features for the re-identification task.

The work of [14] proposes learning a Mahalanobis distance for computing a k-nearest neighbor classification using a maximum margin formulation. Similarly, [15] applies a probabilistic framework for distance function learning. In particular, they search for a distance that maximizes the probability of a matching pair having a smaller distance than a non-matching pair. The very high computational complexity and the strong dependency on the training data are the main drawbacks of these methods. A more efficient and still effective metric learning approach has been proposed in [16]. A preliminary PCA step is proposed to reduce the dimensionality and remove noise; thus, a Mahalanobis metric is learned during the training step.

Despite the performance improvements obtained with metric learning, current re-identification methods still require more powerful descriptors. In the last years, *3D models* have drawn increasing attention in the field of people surveillance. A first re-identification attempt toward 3D body models has been proposed in [17], where local color descriptors are mapped to a cylindrical surface (Panoramic Appearance Map). Recently, also thanks to off-the-shelf sensors and hardware equipment such as the Microsoft Kinect, 3D approaches are becoming more popular. For example, in [18] a set of ratios of joint distances is used as person signature. However, the intrinsic noise on the estimation of the joint positions does not allow us to reach very high performance levels.

To overcome the limits and shortcomings of these body models, we have proposed a new simplified 3D body model. A very preliminary attempt is presented in [19]. In this paper we describe the complete version of the SARC3D model, with a revised and more detailed mathematical formulation. A more complex feature set composed by RGB histograms in a *Mixture of Histograms* formulation for each vertex of the model has been introduced. Finally, a new intermediate step for model alignment to the person's appearance image [20] has been integrated into the system. The complete proposal has been evaluated with a comprehensive set of experiments on the new benchmark suite 3DPeS [21].

3 Exploiting 3D models for people re-identification

In this work we deal with the problem of people re-identification in typical surveillance settings, where multiple cameras, often with disjoint fields of view (FoV), can catch the presence and the appearance of many people walking and crossing a monitored environment.

The main advantages of using 3D models can be better grasped looking at Figure 1. Global features are easy to compute and they don't require alignment steps. In addition, they are more insensitive to segmentation or detection errors. However, the lack of local features leads to deceptive comparisons. The adoption of 2D models partially mitigates the problem, allowing part-based matching. Upper and lower body parts are correctly compared, increasing the specificity of the signature. However, problems mainly related to different orientation may arise. In particular, it is not possible to "look for details". 3D body models, instead, allow us to match the coherence of the person's global appearance as well as the presence/absence of identifying details. For example, the snapshots in Figure 2 can be correctly associated with the same person by both checking the color global correspondence and looking for the white details of the T-shirt.

Appearance-based methods still suffer from color balancing and calibration issues. To this aim, a lot of pre-processing tasks have been introduced to mitigate the related drawbacks. A method for matching appearances using different cameras involves finding a transformation that maps colors in one camera to those in the other cameras. With this in mind, linear algebraic models [22] as well as more complex non-linear approaches [23] have been implemented. When illumination changes are solely responsible for color incoherence among cameras, Brightness Transfer Functions can be

an invaluable tool if learned and applied before the processing, as suggested by Porikli *et al.* [24], Javed *et al.* [25], and Gilbert *et al.* [26]. Due to the high variability of real conditions, the color calibration issue has been neglected in the following and image sources are considered already calibrated.

Fig. 2 Different snapshots of the same pedestrian viewed by a network of cameras, under varying light conditions

The overall schema of our re-identification framework is shown in Figure 3. The frames collected by the network of cameras are first processed using common surveillance systems (left block): people are extracted by means of a people detector [27,28] followed by a segmentation algorithm if needed, or with a people tracker based on background suppression. For example, we extracted the bounding box and the segmentation foreground-background of the 3DPeS shots using a static-camera background subtraction, people detection and tracking described in [29]. Since the number of images available for each person can be very high, a shot selection step is usually required to discard redundant or misleading images. At the end of this stage, a set of K shots $I_1^p \dots I_K^p$ are collected for each candidate person p. Matching of these multiple shots is assured by the tracking algorithm. However, different sets of shots may correspond to the same person and discovering any such relation is the objective of re-identification.

For each one of the selected shots, a single-shot signature is computed through a 3D model alignment followed by an image to model mapping. Descriptors belonging to the same person are integrated into a unique signature, which is compared with a set of stored models to find possible matches. At the end, the new model can be included in the stored archive, if required.

4 The SARC3D Body Model

The reference body model (called *SARC3D*) is based on a vertex set, inspired by the meshes commonly used in computer graphics (see Fig. 4(a)). The model is composed of two parts: a geometry template and an appearance signature. The geometric template is fixed for every person and based on a scale factor only. The vertex positions have been manually defined using a sarcophagus-like shape and exploiting real data to catch some average proportions of the body. Specifically, some side, frontal and top shots of people were extracted from surveillance videos and the average silhouette (Fig. 4(b)) exploited to create a 3D hull (Fig. 4(c)).

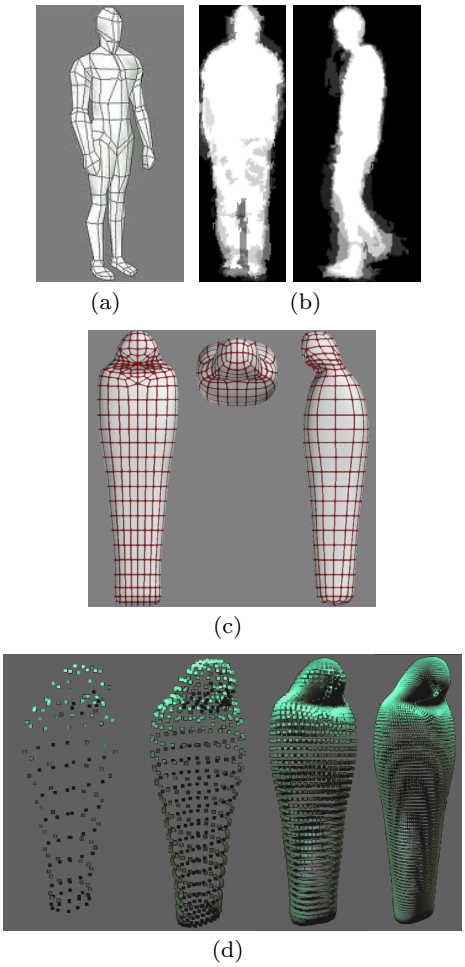


Fig. 4 Genesis of SARC3D: (a) a human 3D model, (b) average silhouettes used for model creation, (c) our simplified human model, and (d) different sampling densities of the SARC3D model used in our tests

The set of vertices

$$\mathbf{v} = \{v_i\}, i = 1 \dots M \tag{1}$$

has been regularly sampled from the sarcophagus surface (Fig. 4(d)). The number M of vertices could be selected accordingly to the required model resolution. Regular sampling assures that each triangle of the mesh has the same area, and therefore the same weight when used in the following computations. In our tests on real surveillance setups, we have sampled the surface obtaining a set of $M = 628$ vertices (Fig. 4(d)). Other

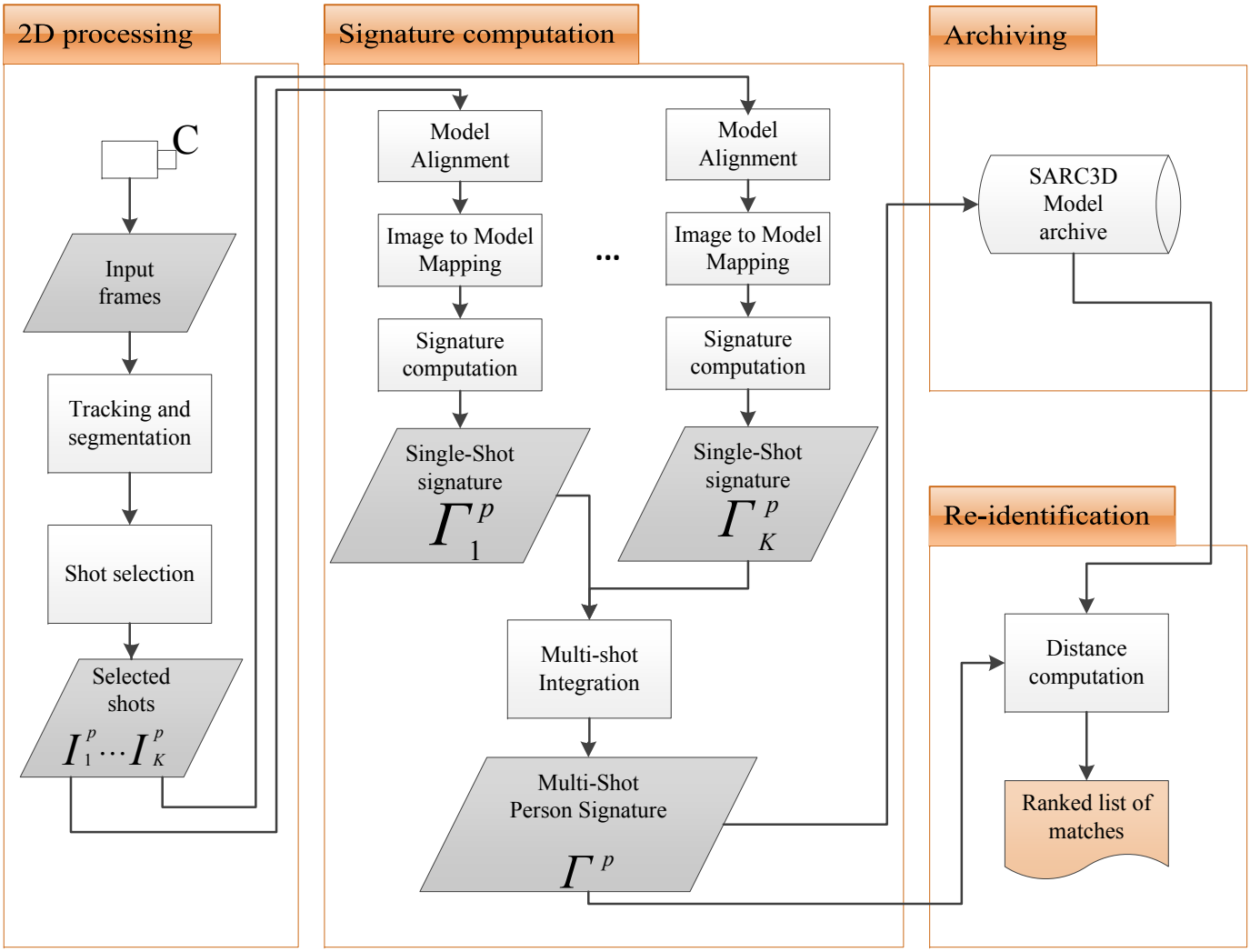


Fig. 3 Overall schema of the re-identification framework

sampling densities were also tested; however, small variations in the number of samples (using N between 500 and 700) do not affect model performance. Larger variations, instead, lead to worse performance due to over- or under-fitting issues as reported in our preliminary work about the SARC3D model [30]. Additionally, the selected value is a good trade-off between speed and efficacy.

The coordinates (X_i, Y_i, Z_i) related to the reference coordinate system (O_M, X_M, Y_M, Z_M) constitutes the final geometric template, as shown in Fig. 5(a). The directions of the vectors $\mathbf{n}_i$ orthogonal to the sarcophagus surface are also computed on each vertex v_i. To cope with different people's heights, the template coordinates have been normalized such that $\max_i Z_i = 1$. The SARC3D model is publicly available on ViSOR [31].

According to 2D methods [10], the vertices have been also divided into 20 horizontal stripes by partitioning them using their Z_M coordinate. A glossary of the main symbols used in the paper is given in Table 1, for the sake of clarity.

4.1 Vertex Feature Set

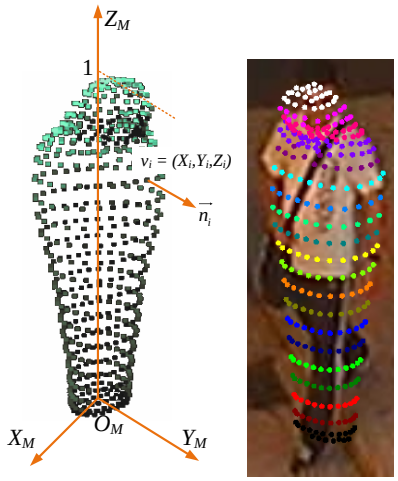
A specific appearance signature is computed on each person shot I^p_l. Rather than providing a bitmap texture for each triangle as in computer graphics approaches, a visual descriptor Γ^p_i is assigned to each vertex v_i of the geometric template. Finally, the set of vertex-wise descriptors will make up the person signature Γ^p:

$$\Gamma^p = \{\Gamma^p_i\}, i = 1 \dots M \tag{2}$$

The descriptor Γ^p_i contains the visual appearance of the person in the 3D position defined by the vertex v_i. The vertex descriptor contains a set of color and

Symbol	Description
I_l^p	l-th shot of the person p
R_l^p	Bounding box around the person detected in I_l^p
(O_I, X_I, Y_I, Z_I)	reference coordinate system centered in the camera optical center. The subscript I is used instead of I_l^p for sake of clarity, even if the coordinate reference may change from shot to shot.
(O_W, X_W, Y_W, Z_W)	world reference coordinate system
(O_M, X_M, Y_M, Z_M)	model reference coordinate system
$\mathbf{P_i}$, $\mathbf{P_e}$	Intrinsic and Extrinsic calibration parameters of the cameras used to acquire the shot I_l^p.
$\mathbf{v} = \{v_i\}$	set of the model vertices of the geometrical template,
M	number of vertices in the 3D model (628 in our experiments)
$Stripe(v_i)$	horizontal stripe $(1 \dots 20)$ containing the vertex v_i
$\mathbf{n}_i$	normal vector of the sarcophagus surface computed in v_i
$\overrightarrow{v_i O_I}$	vector from the vertex v_i to the optical center O_I
h_l^p	real height of the person p in the shot I_l^p
Γ^p	signature of the p-th person
Γ_i^p	vertex descriptor associated with the i-th vertex i of the p-th person
$\mathbf{Hr}$	RGB color histogram
$\mathbf{Hh}$	HSV color histogram
$\mathbf{Hg}$	histogram of the gradient orientations
ρ_i	vertex reliability. $\rho > 0$: vertex directly initialized from the shot; $-1 \leq \rho < 0$: vertex predicted using a symmetry hypothesis; $\rho = 0$: vertex estimated from its neighbors,
$D(\Gamma^p, \Gamma^t)$	distance function between a couple of models p and t.
α, β, γ	linear coefficients using to weight the RGB, HSV and HoG histogram distances during the computation of $D(\Gamma^p, \Gamma^t)$

Table 1 Symbol glossary



(a) (b)

Fig. 5 The SARC3D model. a) the geometric template and b) the vertices are colored based on the horizontal stripe clusters

gradient histograms instead of a single color to correctly generalize the model and avoid over-fitting issues:

$$\Gamma_i^p = \{\mathbf{Hr}_i, \mathbf{Hh}_i, \mathbf{Hg}_i, \rho_i\}, \quad (3)$$

where $\mathbf{Hr}_i$ is the RGB color histogram, $\mathbf{Hh}_i$ is the HSV color histogram, $\mathbf{Hg}_i$ is the histogram of the gradient orientation [27], and ρ_i is the vertex reliability. The person index p and the shot subscript l have been omitted for clarity. The RGB color histograms have been normalized and quantized using 8 bins for each channel; HSV color histograms adopt 8 bins for the H channel, 4 bins for the S and V ones; Histograms of Oriented Gradients (HOG) contain 9 bins. Since hue (H) is more important than saturation (S) and value (V) components [32] in the human visual system, we have assigned more bins to the H channel than to the other components, as proposed in [33] as well. The final dimension of each vertex descriptor Γ_i^p is thus equal to 50, leading to a 31400-dimensional signature Γ^p associated with the person p.

4.2 Image to Model Mapping

Let (O_W, X_W, Y_X, Z_W) be a world reference coordinate system, with Z_W indicating the vertical direction as depicted in Figure 6 (b). Let I_l^p be the shot containing the l-th shot of the person p and R_l^p the corresponding bounding box (see Fig. 8). The position and internal calibration of the acquiring camera are known, i.e., the relations between the image plane π, the camera optical center O_I and the world coordinate reference O_W are defined [34] (see Fig. 6).

Placement of the model into the 3D scene makes Image-to-Model mapping possible. This is defined by the rotation and translation matrices between the world reference coordinate system and the model one (O_M, X_M, Y_M, Z_M)

(see Fig. 6(c)). The last parameter required is the real height of the model h_l^p, which defines the scale factor to apply to the geometric template. The projection (x_i, y_i) of the vertex v_i onto the image plane π is thus defined as follows:

$$\begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \propto \mathbf{P_i} \cdot \mathbf{P_e} \cdot \mathbf{R_l^P} \cdot \mathbf{T_l^P} \cdot \begin{bmatrix} h_l^p \cdot X_i \\ h_l^p \cdot Y_i \\ h_l^p \cdot Z_i \\ 1 \end{bmatrix} \quad (4)$$

where $\mathbf{P_i}$ and $\mathbf{P_e}$ are the perspective projection matrix (intrinsic parameters) and the extrinsic calibration matrix respectively, while $\mathbf{R_l^P}$ and $\mathbf{T_l^P}$ are the rotation and translation matrices from the model coordinate system to the world one, respectively. The algorithm to estimate the placement parameters $\mathbf{R_l^P}$, $\mathbf{T_l^P}$, and h_l^p is described in Section 4.3.

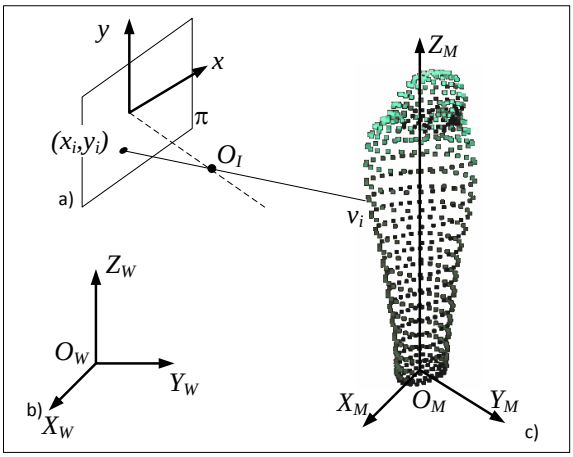


Fig. 6 Schema of the a) Camera, b) World and c) Model reference coordinate systems.

Using Equation 4, each model vertex v_i is projected to a corresponding image point (x_i, y_i). Let R_i be a neighborhood of (x_i, y_i). In our experiments, R_i is a squared region of 9x9 pixels. The uniform sampling used to generate the vertex set guarantees that each descriptor corresponds to the same surface area and thus has the same weight. The color and gradient histograms included in the signature Γ_i^p are then computed using the region R_i as shown in Figure 8.

A reliability value ρ_i is included in the vertex model and takes into account how accurately and precisely the vertex descriptor has been captured from the data. It is computed as

$$\rho_i = \frac{\mathbf{n}_i \cdot \overrightarrow{v_i O_I}}{||\mathbf{n}_i|| \cdot ||\overrightarrow{v_i O_I}||}, \quad (5)$$

where $\overrightarrow{v_i O_I}$ is the unit vector starting from the vertex v_i and directed toward the camera optical center O_I, while $\mathbf{n}_i$ is the vector normal to the 3D body surface (as defined in Section 4).

Figure 7 contains a top-view sketch of the image to model mapping. Two people have been detected in the scene, thus two corresponding models Γ^r and Γ^s have been placed and should be updated. Each vertex v_i is projected onto the (virtual) image plane I. The direction of the normal vector $\mathbf{n}_i$ does not depend on the position of the image plane, while the vector $\overrightarrow{v_i O_I}$ starting from the vertex and directed toward the optical center O_I is related to the person placement and orientation with respect to the camera. The image point (x_i, y_i) and its neighbors are exploited to compute and update the descriptor of the vertex v_i of Γ^r. The cosine of the angle α_i — i.e., the dot product of Equation 5 — is used as vertex reliability.

The reason behind the adoption of the dot product is that data from front-viewed vertices and their surrounding surface are more reliable than those from laterally viewed vertices. This reduces the drawbacks due to errors in model positioning and orientation, since we assign stronger weights to the central points of people and lower weights to lateral points, which are the most hit by misalignment.

For example, a local image region of 9x9 pixels is used for extracting histogram features for each vertex, even if the vertex is close to the boundary of the foreground mask. In these cases, the local image regions will contain background points and the corresponding extracted histograms will be affected by the background. However, reliability will be close to zero since the normal vectors will lead to mark these vertices as lateral.

In addition, vertices hidden in the 2D image (e.g., the back of the right arm in Figure 9(c)) have a dot product with a negative value indicating a lack of reliability. Computer graphic rendering libraries are used to speedup the mapping process and the computation of the dot products included in Equation 5, allowing real time performances to be attained.

A correct model alignment step is critical for obtaining acceptable re-identification performance. Therefore, a precise and accurate bounding box estimation of the person in the 2D shots is desirable. Some quantitative experiments on the capability of the proposed method to face the errors arising during bounding box estimation and segmentation are reported in [35].

Symmetry and neighborhood based prediction When mapping a 2D shot on a 3D model, more than half of the descriptors cannot be set from a single shot, since the corresponding vertices are located in the non-visible part

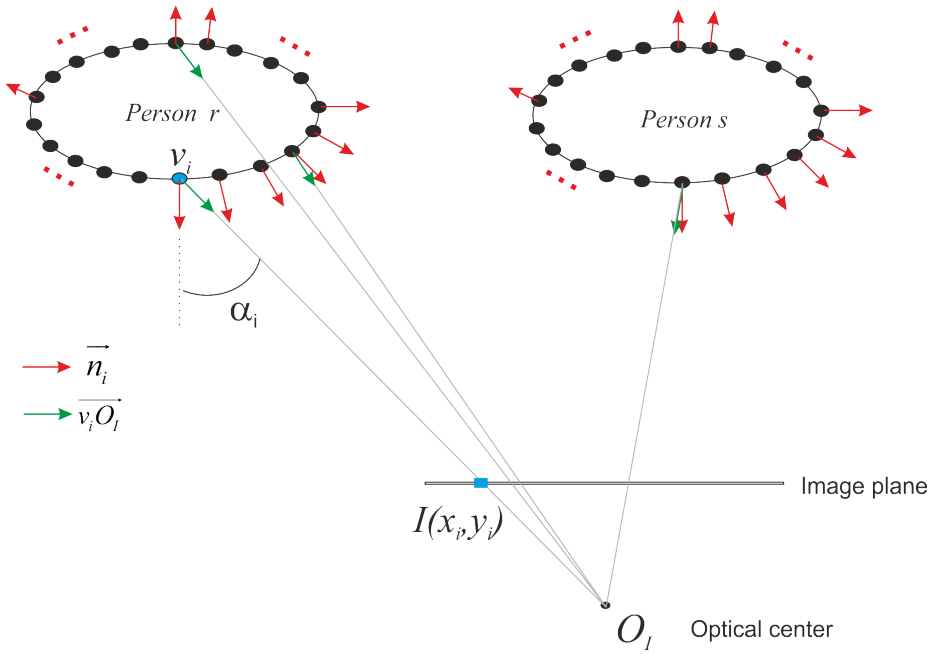


Fig. 7 Sketch of the image-to-model mapping step

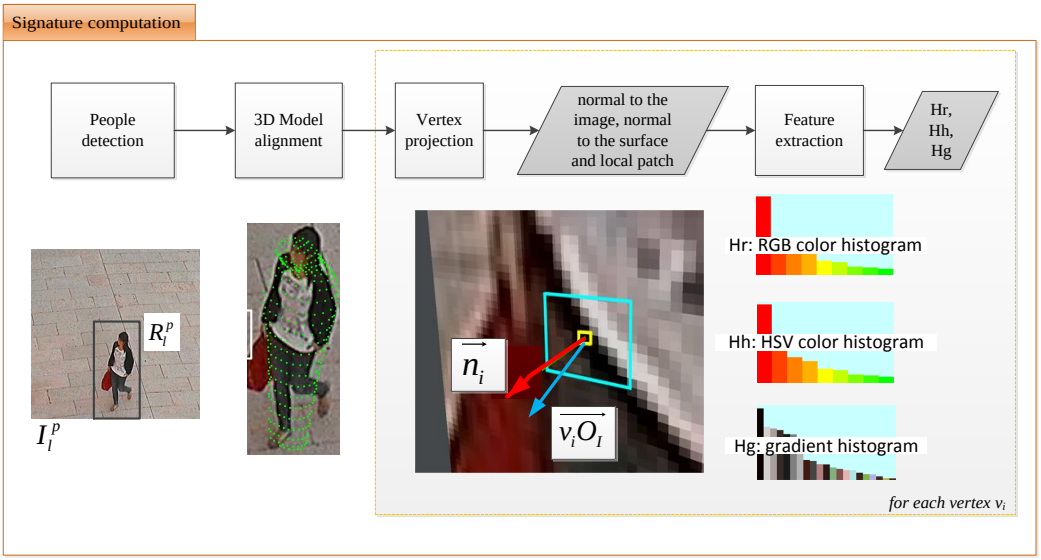


Fig. 8 Extraction of color and texture features

of the model or they are projected outside the image foreground. However, each model vertex is initialized thanks to a prediction step based on symmetries and neighborhood propagation.

In particular, two cases may be outlined:

a) Vertices without a match on the shot; some of the model vertex are projected outside the image foreground, e.g., when the person is not perfectly standing up or when the foreground is affected by segmentation errors. For example, the vertices colored in light blue in the example of Figure 9(b) are projected outside the image foreground. In these cases, the vertex descriptors are estimated from the visible vertices on the neighborhood. To this aim, the vertices have been divided into ten horizontal stripes (see Fig. 9(a), where each stripe is identified by a specific color). The predictions of the vertices without an image match are initialized using the average of the feature vectors belonging to the same stripe, while their reliability values are set to the minimum value (i.e., $\rho_i = 0$).

b) Vertices belonging to the rear (and thus occluded) side of the model; they are also mapped to the image and initialized as the visible ones. However, their reliability will be negative due to the opposite directions

of $\mathbf{n}_i$ and $\overrightarrow{v_i O_I}$. This way, each vertex of the model is initialized even with a single shot: from the real shot, if available, or using a sort of frontal/rear symmetry hypothesis in the absence of information. Even if both the people's body and clothes are not front-rear symmetric, a good re-identification rate can be achieved using the opposite side [36]. This assumption strongly depends on the dataset or the application environment. For example, many people in the ViPER dataset carry a backpack, which invalidates the front-rear symmetry assumption. We used the term "prediction" in order to highlight its aleatory foundation. The predictions are used in the absence of information only, and they are immediately replaced by real data whenever these are available. However, if the dataset or the working conditions discourage the use of this assumption, the rear vertices can be left empty or set with the stripe prediction previously described (case a).

By means of the reliability value, vertices directly initialized from the image ($\rho > 0$), vertices predicted using a symmetry hypothesis ($-1 \leq \rho < 0$) and vertices estimated from their neighbors ($\rho = 0$) are distinguishable.

The steps of the initialization phase are described in the pseudo-algorithm 1. The direct computation of the Γ descriptor (line 7) is carried out for both frontal and rear vertices. In the first case, the descriptor is also used to generate the stripe descriptor (lines 11 and 12) which will be assigned to all the vertices outside the image bounding box or the foreground silhouette (lines 20-23).

Algorithm 1 Image to Model Mapping

Require: $\mathbf{P_i}, \mathbf{P_e}$ ▷ Calibration parameters
Require: $\mathbf{R_l^P}, \mathbf{T_l^P}, h_l^p$ ▷ Model placement
Require: I^p, R^p ▷ Input shot and bounding box of the person

Direct feature computation
1: **for all** $v_i = (X_i, Y_i, Z_i) \in \Gamma_i$ **do**
2: $(x_i, y_i) \leftarrow Projection(v_i)$ ▷ Projection of the vertex on π
3: $R_i \leftarrow N(x_i, y_i)$ ▷ Neighborhood of (x_i, y_i)
4: **if** $R_i \cap R^p = \emptyset$ **then**
5: $\rho_i = 0$ ▷ v_i will be predicted using Stripes
6: **else**
7: Compute Γ_i on R_i
8: $\rho_i \leftarrow \mathbf{n}_i \cdot \overrightarrow{v_i O_I}$
9: **if** $\rho_i > 0$ **then**
10: v_i is a frontal point
11: $j = Stripe(v_i)$
12: Update $\bar{\Gamma}_j$ with Γ_i ▷ Stripe mean computation
13: **else if** $\rho_i < 0$ **then**
14: v_i is a rear point ▷ Γ_i is the Symmetry based prediction
15: **else**
16: ▷ v_i will be predicted using Stripes
17: **end if**
18: **end if**
19: **end for**

Stripe based prediction
20: **for all** $v_i = (X_i, Y_i, Z_i)$ such that $\rho_i = 0$ **do**
21: $j = Stripe(v_i)$
22: $\Gamma_i = \bar{\Gamma}_j$
23: **end for**

4.3 Model Alignment

The alignment of the 3D model on each shot is one of the most critical steps within this approach. More precisely, the person height h_l^p and the Rotation and Translation matrices $\mathbf{R_l^P}$ and $\mathbf{T_l^P}$ defined in Section 4.2 should be estimated given a shot I_l^p. To simplify and speed up the process, we impose a constraint on the person posture, which is assumed to be standing.

The matrix $\mathbf{T_l^P}$ contains the three coordinates of the person with respect to the world reference coordinate system. Assuming that the person is standing on the ground plane, a homography transformation of the feet position (lower point of the vertical axis of the bounding box) [37] can solve the placement step.

The rotation matrix $\mathbf{R_l^P}$ has only one degree of freedom, which is the angle θ_l^p around the vertical axis Z_M. The other two rotation matrices are defined by assuming the standing posture constraint. The estimation of θ_l^p is detailed in the following.

The height h_l^p is finally obtained by assuming that the projected model vertices lie and fill the bounding box R_l^p of the detected person in the image plane. In practice, the height is estimated starting from the height h of the bounding box R_l^p which contains the person foreground and using the inverse projection matrix. A visual example of the model placement and orientation estimation is depicted in Figure 10.

Orientation estimation If the detection in a shot I_l^p comes from a video sequence and the person trajectory has been provided through a tracking algorithm, the orientation can be inferred from the trajectory itself. Otherwise, the person orientation is estimated on still images exploiting the method proposed in [20]. In the first case, the angle θ_l^p is computed by averaging the direction and the orientation of the last frames. This approach is simple and without computational burden if tracking is available, but being a prediction it is subject to frequent failures since people may change direction abruptly and, in addition, it may not be applied to still people where their orientation does not depend on trajectory.

In the second case, we propose a supervised orientation classification of appearance based on HOG descrip-

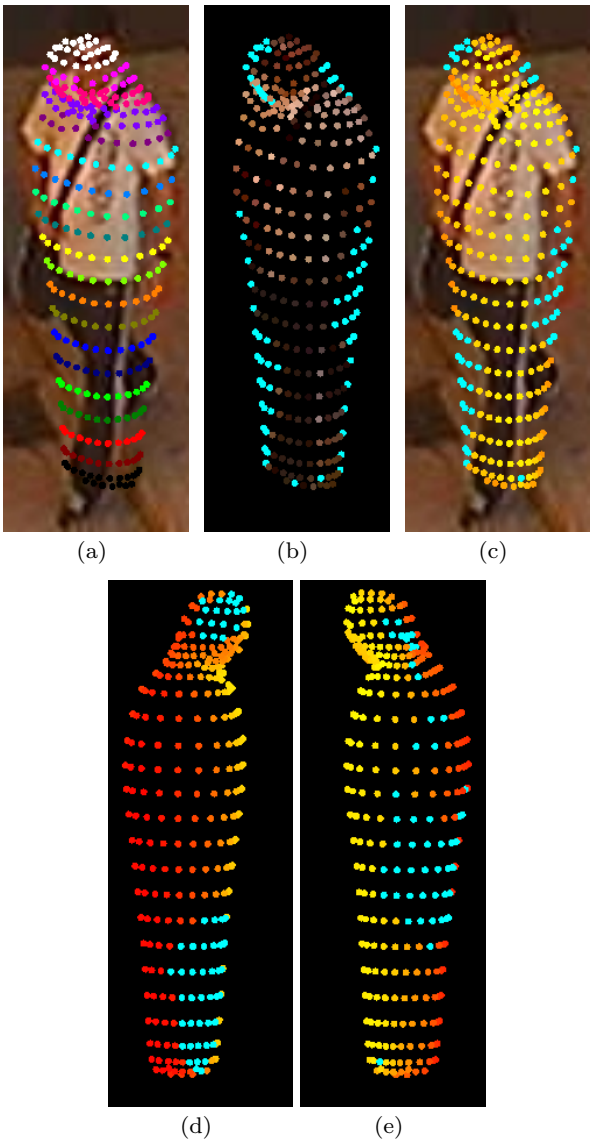


Fig. 9 (a) SARC3D projection, (b) Color feature extraction, (c) Vertex reliability, (d,e) Left and right views of the vertex reliability

tors. Angles θ_l^p are firstly quantized in $N = 8$ discrete values $\theta^i \in \left[\pi + k\frac{\pi}{4}, k = 0 \dots 7\right]$. A set of N Extremely Randomized Forest classifiers are thus trained with a suitable labeled set.

Each classifier provides a score value which is integrated into a global continuous probability density function using a Mixture of Approximated Wrapped Gaussian distributions in order to estimate the correct continuous orientation θ_l^p.

For each detected person, a 2268-dimensional feature vector is loosely computed based on the HOG descriptor[27]. The appearance image of the person is first converted into a gray-level image. The image is then split into blocks at three different levels: the first

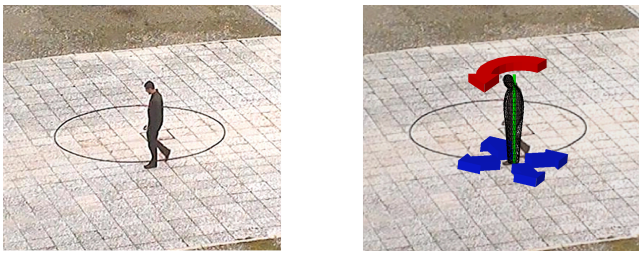


Fig. 10 Model placement (blue arrows) and orientation (red arrows) of the 3D model on top of the input frame (left image)

level contains 8x24 non-overlapping blocks, the second level 4x12 blocks and the last level 2x6 blocks. At each level, the image is down-sampled with a scale factor of 0.5. A histogram of oriented gradients quantized in 9 wrapped bins is computed on each of the 252 blocks using tri-linear interpolation and normalized over sets of four blocks. The 9 histogram values of the 252 blocks are concatenated, obtaining a 2268-dimensional feature vector which acts as the appearance descriptor of the images and is sent to an ensemble of classifiers.

The Extremely Randomized Trees classifiers introduced by Geurts *et. al.*[38] have been adopted due to the very high dimensionality of the input feature vector and since they return a classification score ψ^i instead of a binary response. Formally, given a feature vector ϕ_l, each of the N classifiers provides a value $\{\psi^i \in [0,1], i = 1, \dots, N\}$. A first orientation estimation $\bar{\theta_l}$ of the person in the shot I_l^p could be directly obtained from the corresponding outputs $\Psi_k = \left\{\psi_k^1 \dots \psi_k^N\right\}$ of the classifiers:

$$\bar{\theta_l} = \theta^j, j = \underset{i}{\operatorname{argmax}}\, \psi_l^i. \tag{6}$$

However, the orientation classes previously defined slightly overlap each other. In fact, each class is closely related to its neighbors and the opposing class, which results in more than one high response from the set of discrete-orientation classifiers. Rather than directly using the outputs of the N trained classifiers to generate a discrete class label, the integration of the results of the N classifiers into a continuous probabilistic distribution $p(\theta|I)$ allows us to filter out errors in the orientation estimation as well as to obtain a more accurate value. To this aim, the fuzzy-predicted class label terms ψ^i are used as weights of a mixture of Approximated Wrapped Gaussian distributions [39], each centered on the N selected orientations θ^i.

A mixture of $\mathcal{AWN}$ is obtained as a weighted sum of $\mathcal{AWN}$ probability density functions:

$$Mo\mathcal{AWN}(\theta|\boldsymbol{w}, \boldsymbol{\theta}_0, \boldsymbol{\sigma}) = \sum_{i=1}^{N} w_i \cdot \mathcal{AWN}(\theta|\theta_{0,i}, \sigma_i), \tag{7}$$

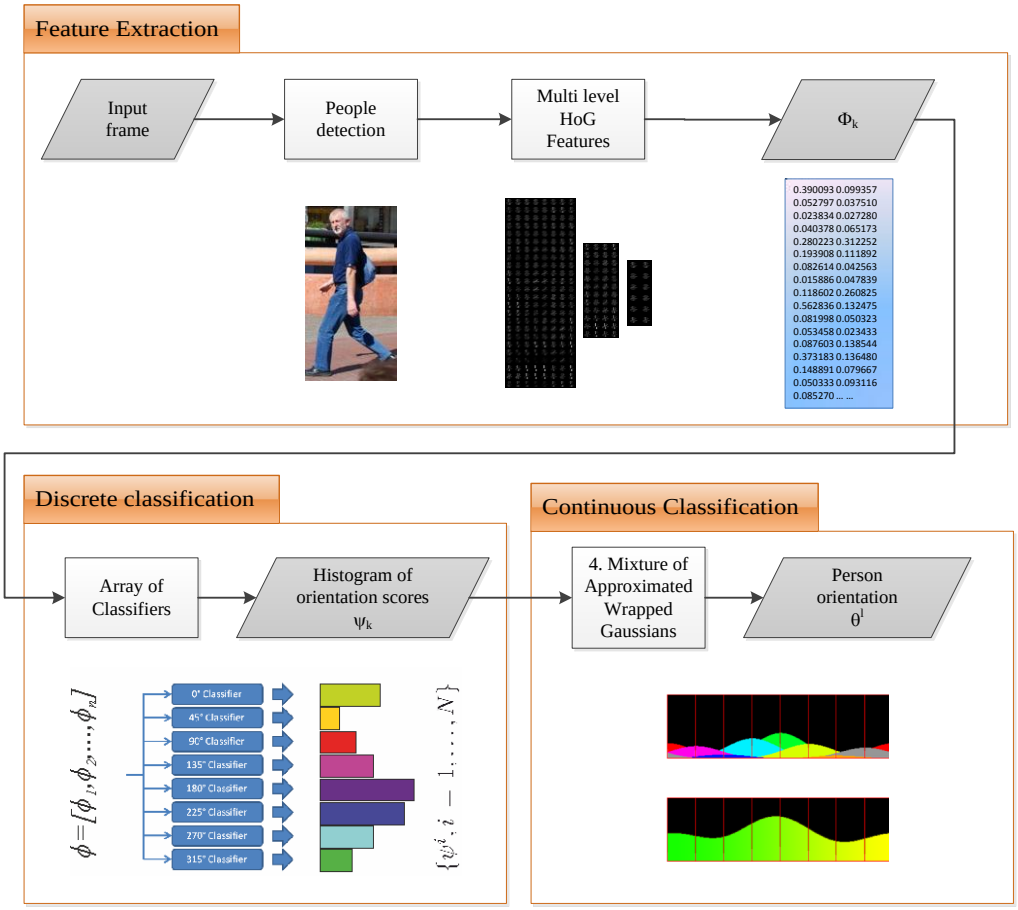


Fig. 11 Schema of the orientation estimation algorithm: Input shot, Multi-Level HOG, Array of classifiers, the Mixture of Approximated Wrapped Gaussians.

where

$$\mathcal{AWN}(\theta|\theta_0,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{((\theta-\theta_0) \mod 2\pi)^2}{2\sigma^2}}. \quad (8)$$

The required function for the orientation estimation is thus obtained using a Mixture of Approximated Wrapped Gaussian as in Equation 7:

$$p(\theta|I_l) = Mo\mathcal{AWN}(\theta|\psi_l, \boldsymbol{C}, \sigma), \quad (9)$$

where the variance σ has been set to a fixed value for all the components.

The orientation θ_l^p is then computed by maximizing the previous distribution:

$$\theta_l^p = \underset{\theta\in[-\pi\ldots\pi]}{\operatorname{argmax}}\ p(\theta|I_l). \quad (10)$$

through Mean-Shift optimization. To avoid estimating a local minimum, the Mean-Shift procedure is repeated starting from the eight discrete values θ^i as seeds. Figure 11 shows all the steps of the proposed method and, in particular, the final filtering step obtained with the Mixture of Approximated Wrapped Gaussians.

Fig. 12 Some graphical results of the model orientation and positioning on the 3DPeS dataset

The estimation of people orientation is an intrinsically continuous problem. The discretized classes are not clearly separated and sometimes are even overlapped (due to the torsion movements of the body). Taking into account the output of different trained classifiers instead of a unique response allows us to better cope with the problem. Each binary classifier is trained to select a positive region of the feature space, which is not related and not imperatively disjoint with the others. The additional step we propose integrates different contributions and is therefore capable of improving the final classification when orientation is quite ambiguous.

The σ parameter of Equation 9 depends on the number of adopted classifiers: if σ is 0 the AWG step is disabled. Increasing the σ value allows us to include the contributions of more neighbor classifiers in the final response. In the experiments, we used $\sigma = 0.75$ in order to integrate also the contribution of the two neighbor classifiers into each main direction.

The final orientation classifier has been trained using images from the TUD Multiview Pedestrian dataset [40] and SARC3D dataset [19].

5 Multi-shot integration

The capability of integrating multiple shots is one of the major advantages provided by the adoption of 3D body models. As described in Section 2, many recent re-identification methods exploit multiple shots, often integrating them into a single multidimensional descriptors or averaging their values. Using the SARC3D model the integration of multiple shots can be handled at vertex level.

5.1 Shot selection

Especially in surveillance applications, multiple shots of the same person are automatically extracted from a video sequence using a tracking algorithm. In this case, redundant or noisy shots can negatively affect re-identification performance, both in terms of computational complexity and error rate. In such cases, a shot selection step is required to keep the most important items for the model initialization and updating. We propose here a rule-based approach to select the best shots in surveillance applications. Selection checks can be of two different types: *shot rules* (i.e., the entire shot is selected or neglected) or *vertex rules* (i.e., the check and the related decision are applied to each vertex, independently). Similar or additional rules can be included depending on the specific application or implementation.

5.1.1 Shot rules

Occlusion check: after the model placement and alignment (see Section 4.3), each couple of models is analyzed to detect occlusions. To avoid false pixel-to-model assignments, both the occluding and the occluded models are not updated. A visual example of the 3D occlusion detection is shown in Figure 13.

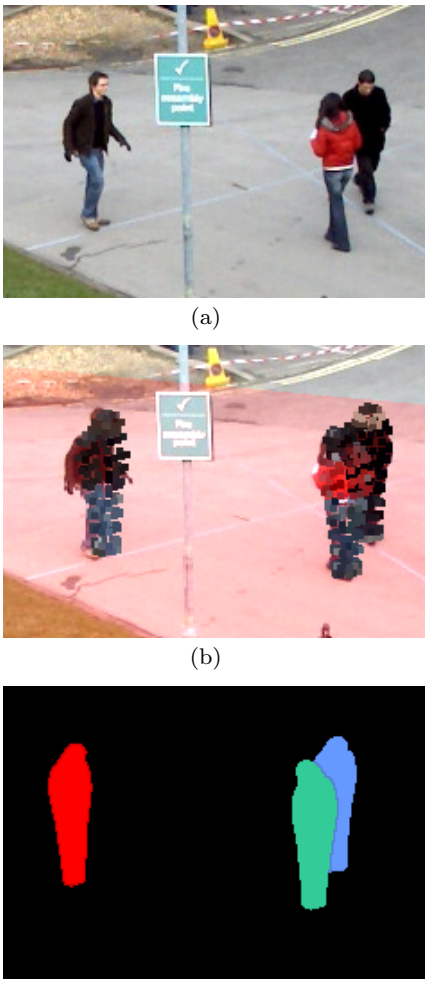


Fig. 13 Occlusion detection: (a) the input frame, (b) the aligned 3D models and (c) the masks $\hat{I}_p$ generated by the rendering system. Since the blue and green objects are connected, the corresponding models are frozen and not updated during occlusion

Model to foreground overlapping: for each shot, an overlapping score R_p is computed as the ratio between the number of foreground pixels that overlap with the projected masks of the model vertices into the image plane ($\hat{I}_p$, see Fig. 13(c)) and the total number of foreground pixels. If R_p is higher than a strong threshold (e.g., 95% in our experiments) the selected shot is marked as good. If the alignment is not precise enough or the person is not in a posture compliant with the sarcophagus model, the R_p decreases and the shot is rejected.

In the case of Figure 13, the detection-by-motion system extracts people silhouettes as foreground regions. During the occlusion with the pole, the corresponding region will be excluded from the foreground mask and the corresponding hidden vertices will remain without description or will be associated with a pre-

dicted one. If the occlusion by the pole is limited and included in the foreground, the pole appearance will be used to update some of the model vertices. In this case, only few vertices will be affected by errors and the integration of additional shots will probably solve the problem. On the other hand, if the occlusion by the pole is too heavy, the model to foreground overlapping check will completely disable model updating.

The 3D model can be updated even if the 2D processing part provides a bounding box for each person instead of a precise foreground silhouette (see Fig. 3). In this case, the model update is still valid thanks to the correspondence between the model vertices and the image points, even if some descriptors will be degraded when the corresponding vertices fall outside the foreground mask (as in the examples of Fig. 12). The Model to foreground overlapping check is not feasible in this situation.

Orientation reliability: the reliability of the orientation estimation may be evaluated considering the sequence of the estimated orientations: if the distribution of the differences between consecutive orientations has a high variance, the trajectory is not stable and the orientation becomes unreliable.

5.1.2 Vertex rules

Information gain: the performance of the mixture of histogram approach decreases if a lot of shots are integrated into the same signature, due to computational issues and storage requirements. Each time a new shot is included in the signature, the new vertex-level histogram is compared with the items already stored in the mixture. If a similar histogram is found (distance defined in Equation 17 under a given threshold), the new vertex histogram is skipped. The same rule can be applied in the averaging histogram approach, even if the computational load of the histogram update (average operation) is similar to the evaluation of the distance of Equation 17 and the storage requirements are not modified.

Visible vs hidden: descriptors with negative reliability are discarded if at least one descriptor of the same vertex has a positive reliability value. In other words, if a vertex has been seen at least once from a frontal point of view, all the predicted descriptors are discarded.

5.2 Signature Combination

For each available shot I_l^p of the same person p, a single-shot signature $\Gamma^{p,l}, l = 1 \dots L^p$ is created as described in Section 4. The obtained items can be integrated into a unique signature $\Gamma_i^p = \{\mathbf{Hr}_i^p, \mathbf{Hh}_i^p, \mathbf{Hg}_i^p, \rho_i^p\}$ by combining each corresponding vertex:

$$\Gamma_i^p = \bigcup_{l=1\dots L^p} \Gamma_i^{p,l} = \left\{ \bigcup \mathbf{Hr}_i^{p,l}, \bigcup \mathbf{Hh}_i^{p,l}, \bigcup \mathbf{Hg}_i^{p,l}, \bigcup \rho_i^{p,l} \right\}. \tag{11}$$

The meaning of the union operator $\bigcup$ applied to each set of corresponding vertex descriptors depends on the selected integration method. Two different approaches have been implemented and tested: *histogram averaging* and *mixture of histograms*. In the first case, the resulting signature is a sort of average of the input items and has the same dimension of the starting ones. In the second case, instead, all the input signatures are embedded in the same container, increasing the signature size. As confirmed by the experiments in Section 7, the integration by averaging has worse performance but allows a more compact representation; its use is recommended if the signature size should be limited due to storage or transmission constraints.

In the mixture approach, the new descriptor will be composed of the set of single-shot descriptors, each one weighted by the original reliability value. The merged RGB descriptor is the set of all the RGB histograms and the corresponding reliabilities:

$$\mathbf{Hr}_i^p = \bigcup \mathbf{Hr}_i^{p,l} = \left\{ \rho_i^{p,l}, \mathbf{Hr}_i^{p,l} \right\}_{l=1\dots L_N}. \tag{12}$$

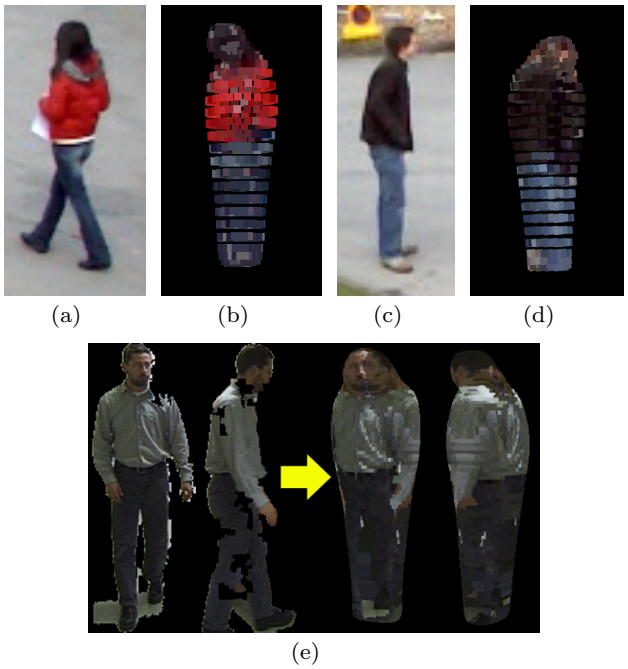
Similar equations are defined for the HSV and gradient histograms.

The model averaging approach, instead, aims at obtaining a more compact descriptor with a linear combination of the single-shot features; a function $\phi(\cdot)$ of the vertex reliabilities $\rho^{p,l}$ is used as weight:

$$\mathbf{Hr}_i^p = \bigcup \mathbf{Hr}_i^{p,l} = \frac{1}{\sum \phi(\rho^{p,l})} \cdot \sum_{l=1\dots L} \phi(\rho^{p,l}) \cdot \mathbf{Hr}_i^{p,l}. \tag{13}$$

Sum, product and division operators of Equation 12 and Equation 13 should be interpreted as applied to each histogram bin. The integration of the other descriptors follows the same methods. Finally, the function $\phi(\rho_i^l)$ allows us to correctly handle both positive and negative values of the reliability. A possible implementation is reported in Equation 14:

$$\phi(\rho) = \frac{1+\rho}{2}, \tag{14}$$



(a) (b) (c) (d) (e)

Fig. 14 Sample 3D models (b,d) created with single shots (a,c) or multiple shots (e).

where the reliability values have been linearly mapped to the $[0, 1]$ interval. Figure 14 shows some examples generated by integrating multiple shots through averaging of the feature vectors.

6 Distance metric for People re-identification

One of the strengths of model based approaches consists in simplifying the matching phase. Given two person signatures Γ^p and Γ^q, their distance D can be decomposed into a sum of vertex-wise distances:

$$D(\Gamma^p, \Gamma^q) = \frac{\sum_{i=1\ldots M} \left(w_i \cdot d(\Gamma_i^p, \Gamma_i^q)\right)}{\sum_{i=1\ldots M} \left(w_i\right)}, \tag{15}$$

where the product of the vertex reliability values is used as weight w_i in order to emphasize the parts of the models which have been seen from a frontal direction:

$$w_i = \left|\phi(\rho_i^p) \cdot \phi(\rho_j^q)\right|. \tag{16}$$

The vertex distance $d(\Gamma_i^p, \Gamma_i^q)$, in turn, is split into three contributions, one for each histogram type included in the vertex descriptor:

$$\begin{aligned} d\left(\Gamma_i^p, \Gamma_i^q\right) = \alpha \cdot d\left(\mathbf{Hr}_i^p, \mathbf{Hr}_i^q\right) + \\ +\beta \cdot d\left(\mathbf{Hh}_i^p, \mathbf{Hh}_i^q\right) + \gamma \cdot d\left(\mathbf{Hg}_i^p, \mathbf{Hg}_i^q\right). \end{aligned} \tag{17}$$

where α, β, and γ are three coefficients which allow the desired feature combination to be selected. Given

Fig. 15 Sample images from the 3DPeS dataset

the number of available shots, called L_p and L_q, embedded in the two signatures Γ^p and Γ^q respectively, the three distances $d(\cdot, \cdot)$ of Equation 17 are defined as the minimum distance between each possible couple of histograms:

$$d\left(\mathbf{Hr}_i^p, \mathbf{Hr}_i^q\right) = \min_{\substack{r=1\ldots L^p \\ s=1\ldots L^q}} d_H\left(\mathbf{Hr}_i^{p,r}, \mathbf{Hr}_i^{q,s}\right), \tag{18}$$

where d_H is the Hellinger distance between histograms [41]. Since illumination or view changes have different effects on the RGB, HSV or HOG histograms, each term in Equation 17 is derived from a different couple of shots, i.e., the indexes r and s in Equation 18 may be different for each histogram type.

7 Experimental results

The proposed method has been extensively tested using the publicly available dataset **3DPeS** (3D People Surveillance Dataset) [21]. 3DPeS is a surveillance dataset, mainly designed for people re-identification in multi-camera systems with non-overlapped field of views (see Fig. 15 for some sample images). Versus other re-identification datasets, such as Viper [42] or ETZH [43], 3DPeS contains short video sequences instead of still images, allowing a complete evaluation of all the system capabilities. Most of the people in the dataset satisfy both the stripe and the frontal-rear symmetry assumptions described in Section 4.2. Thus, both the corresponding prediction algorithms have been enabled during the following tests. A detailed description of all the datasets available for people re-identification has been reported in [44]. Among others, some datasets extracted from i-LIDS [45] fulfill the requirements imposed by SARC3D approach and have been used to test tracking as well as re-identification systems [46]. In this paper, we have only presented quantitative results on 3DPeS, which is freely available.

A collection of snapshots has been extracted for each person, using the shot selection approach described in Section 5.1. All the sequences which do not contain at least three different shots of the person concerned have been excluded from the set. The final set is composed of 1012 snapshots of 192 different people (available on the 3DPeS website).

For each experiment described in the following, we have generated ten random splits of the snapshots between training and query items. The results provided have been obtained by averaging the ten corresponding outcomes.

Being well-established in re-identification and recognition tasks, for each *query* item we have ranked the *training* gallery elements using the distance metric defined in Section 6. The performance results summarized are reported using the cumulative matching characteristic (CMC) curve, which is analogous to the ROC curve for detection problems [42].

The first set of experiments have been devoted to testing different combinations of features, metrics and parameters. Then, a comparison with two state-of-the-art methods has been provided. Each experiment is shown using CMC curves computed by averaging several test runs on random splits of the entire collection into testing and training sets.

7.1 Feature selection and parameter tuning

We first compared the system performance using different combinations of the distance parameters (Equation 17) and the two strategies for multiple shots integration (see Section 5). In particular, the following cases have been considered:

- RGB Average: $\alpha = 1, \beta = 0, \gamma = 0$; integration with model averaging;
- HSV Average: $\alpha = 0, \beta = 1, \gamma = 0$; integration with model averaging;
- RGB Mixture: $\alpha = 1, \beta = 0, \gamma = 0$; integration with mixture of histograms;
- HSV Mixture: $\alpha = 0, \beta = 1, \gamma = 0$; integration with mixture of histograms;
- RGB + HOG Average: $\alpha = 0.5, \beta = 0, \gamma = 0.5$; integration with model averaging;
- RGB + HOG Mixture: $\alpha = 0.5, \beta = 0, \gamma = 0.5$; integration with mixture of histograms;

We selected three shots for the creation of both query and training models. Figure 16(a) shows the corresponding CMC curves. The multi-shot integration using the Mixture approach has always outperformed the median alternatives, as expected. However, the decrease in performance by averaging the models is tolerable, if mandatory network or storage constraints require its use.

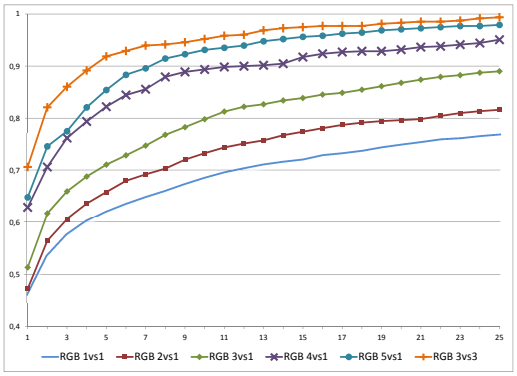


Fig. 17 CMC curve showing the multi-shot integration capabilities. The curves are obtained using RGB features only in the mixture configuration

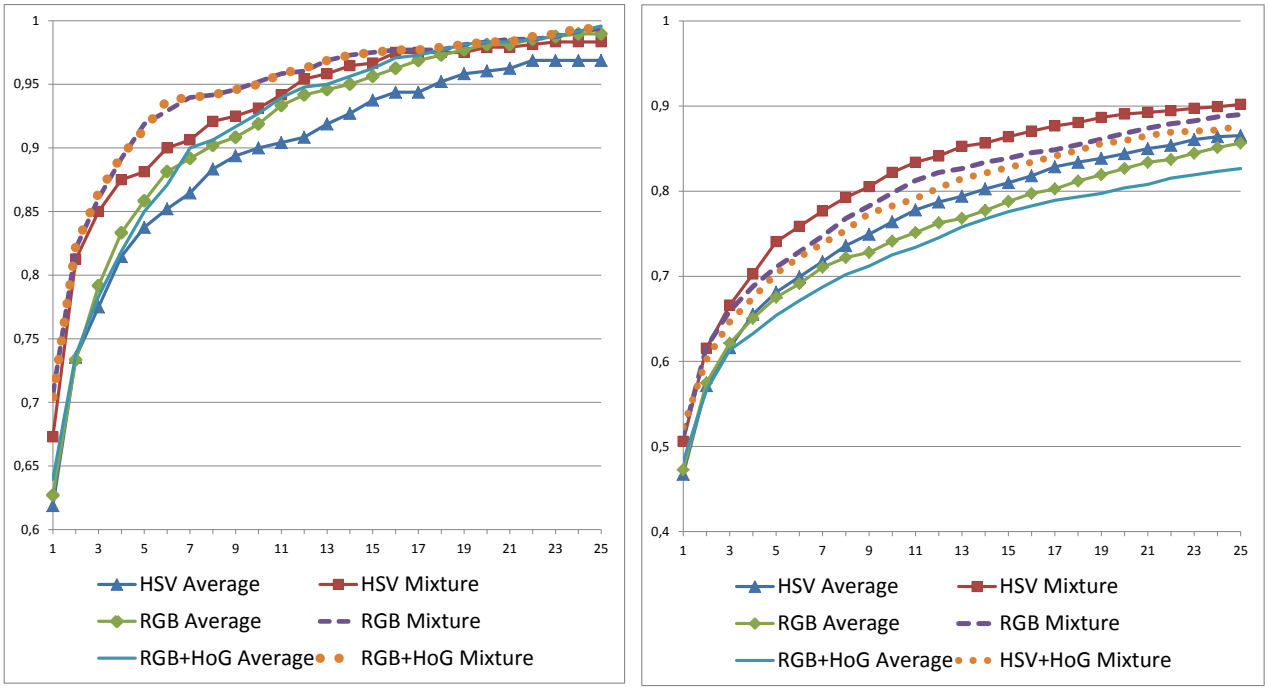
To test the model performance in unbalanced cases, i.e., when the query and the training models are generated by a different number of shots, we have created the training set models using three shots as in the previous experiment, while the query model has been created from a single shot (indicated as *3vs1* in the following). Figure 16(b) shows the corresponding results.

Table 2 shows the AUC (Area Under Curve) measure for the six combinations mentioned above. The first row shows the tests performed using three shots for both query and training models (*3vs3*), the second row reports the tests performed using one shot for the query model and three shots for the training one (*3vs1*). The HOG feature barely improves performance: this is due to the presence of many low-resolution images in the dataset. As further confirmation, we performed a specific test on a subset of high-resolution images (more than $150x50$ pixels). In this case, the HOG contribution improves the system performance by around 5%. The resulting AUCs for the *3vs1* and *3vs3* test cases are 0.971 and 0.983, while the corresponding values using RGB features only are 0.954 and 0.976.

Finally, we evaluated the actual capability of the multiple shot integration task. The improvement in re-identification performances obtained by adding more shots to each SARC3D model is reported in Figure 17. The results are also summarized in Table 3 using the AUC score. The results confirm the multi-shot integration capabilities of the system, especially when the Mixture approach is adopted.

	HSV Average	HSV Mixture	RGB Average	RGB Mixture	RGB+HOG Average	RGB+HOG Mixture
3vs3	0.961	0.971	0.967	0.975	0.968	**0.977**
3vs1	0.889	0.914	0.882	0.905	0.872	**0.898**

Table 2 AUC for the different feature tested: 3vs3 and 3vs1 tests



(a) (b)

Fig. 16 Test on different feature combinations: a) query and test models created with 3 shots each, b) query models created with 3 shots, test models created with one shot only

	1vs1	2vs1	3vs1	4vs1	5vs1	3vs3
RGB+Averaging	0.812	0.852	0.882	0.942	0.958	0.967
RGB+Mixture	0.817	0.856	0.905	0.954	0.967	0.975

Table 3 Mean AUC using models created from an increasing number of shots

7.2 Comparison with the state of the art

We compared SARC3D with three well known re-identification approaches on the 3DPeS dataset. Specifically, we exploited the publicly available implementation of the SDALF method developed by the same authors [12], our implementation of the ensemble of features proposed by Gray *et al.* [47], and our implementation of the Panoramic Appearance Map descriptor [17,48].

SDALF [12] is a purely 2D method. It consists in extracting features that model three complementary aspects of human appearance: the overall chromatic content (using weighted HSV histograms), the spatial arrangement of colors into stable regions (Maximally Stable Color Regions), and the presence of recurrent local motifs with high entropy. All these features are derived from different body parts, and opportunely weighted by exploiting symmetry and asymmetry perceptual principles (each appearance image is segmented into legs/torso/head using simple heuristics).

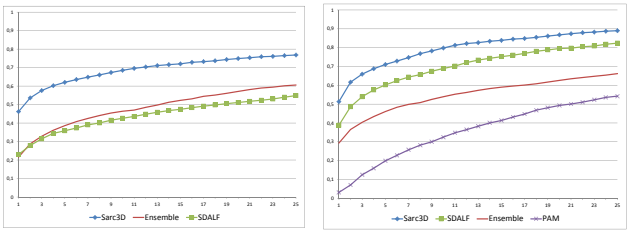
The method proposed in [47] consists of an ensemble of features: RGB, HSV and YCbCr histograms (each channel quantized into 16 bins) and the histograms of the response to 13 Schmid and 8 Gabor filters (each response quantized into 16 bins). The different features are concatenated in a single 464 dimensional feature vector, and a feature vector is computed for each of the 3 fixed size stripes of the person silhouette (roughly head, torso and legs). This method has been extended to the multi-shot case by exploiting a mixture approach. In order to obtain a fair comparison, the distance between feature vectors has been computed through the Euclidean distance between the descriptors rather than

applying an additional metric learning as proposed in the original paper.

Finally, the Panoramic Appearance Map (PAM) descriptor [17,48] is a 3D body model-based approach. A cylindrical model is used as support to map appearance descriptors (mean RGB colors in the paper and in our implementation). The surface of the model is split into a grid of N by M cells ($N = 24$ and $M = 10$). For each cell, together with a mean color, the descriptor contains the number of pixels mapped to the cell, i.e., the number of points used to compute the mean color. This value has been used as weight during distance computation. In the original paper [17], the descriptor were created from multiple shots collected at the same instant and integrated thanks to the 3D camera calibration. People orientation was neglected during computation of the descriptors but was then retrieved when estimating the distance matrix (all the possible alignments between the training and the query models have been explored). In the SARC3D dataset, instead, the different shots have not been captured simultaneously. Thus, the orientation estimation described in Section 4.3 has been used to align the cylinders and to correctly integrate multiple shots.

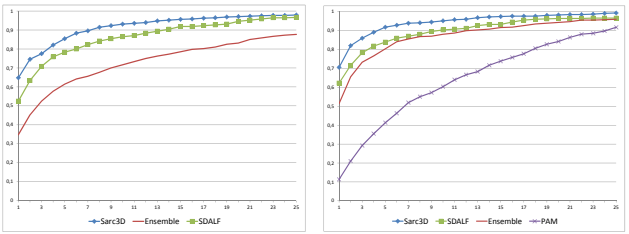
Figures 18 and 19 show different comparisons between the aforementioned methods using different number of shots for creating the training and query models. In detail, Figure 18(a) shows the case in which a single shot is used for both training and query model (*1vs1*). Figure 18(b) depicts the results obtained when 3 shots are used for training model and 1 for test model (*3vs1*). Eventually, Figure 19(a) reports the performance achieved with 5 shots used for training model (*5vs1*), while Figure 19(b) shows performance in the case of 3 shots used for both training and query model (*3vs3*). Table 4 reports re-identification accuracy sampled from the corresponding CMC curves at ranks 1, 5, 10 and 25.

As it can be inferred from the results presented, multi-shot approaches always outperform single shot approaches. However, adding spatial 3D information for feature localization greatly improves re-identification performances as highlighted in Table 4. It should also be noted that the SARC3D algorithm does not require precise model alignment and orientation estimation. As depicted in Figures 12 and 15, the position and orientation estimation is far from being perfect, reaching 61% accuracy on the 3DPeS dataset as reported in [20]. At the same time, the silhouettes extracted using automatic techniques are usually very noisy and contain lots of holes and missing parts as shown in Figure 15. Nevertheless, since each vertex contains a descriptor extracted from a local patch and not from a single image pixel, the system performance is still very promising even in the presence of these issues. Among the approaches tested, the PAM descriptor delivered the worst performances: a mean color for each cell is not enough to catch differences on the texture, which is one of the challenges of the adopted dataset. In addition, PAM seems to be more sensitive to the lack of information; the availability of more than one shot both for the training and the query model dramatically improves the method performance.



(a) (b)

Fig. 18 Comparisons with the state of the art: a)single shot - *1vs1*, b) multi-shot *3vs1*



(a) (b)

Fig. 19 Comparisons with the state of the art: a) multi-shot *5vs1*, b) multi-shot *3vs3*

8 Conclusions

We have proposed a new and effective method for people re-identification. As opposed to currently available solutions, we have exploited a 3D body model to spatially localize the identifying patterns and colors on the model vertices. This way, occlusion and view dependencies are intrinsically solved.

The 3D model proposed is non-articulated and has a fixed shape. As a consequence, the overall system can be correctly applied to images or videos with standing people. If the monitored people are not standing or their shape differs to much from the SARC3D model (e.g., due to backpacks and trolleys as in the i-LIDS dataset [45]), the overall re-identification performance decreases. The alignment between the 3D model and the available images or videos is the most delicate phase

ID	N. of shots		Rank	**SARC3D**	SDALF	Ensemble	PAM
	Training	Query					
1vs1	1	1	1	**46%**	23%	21%	
			5	**62%**	35%	38%	
			10	**68%**	42%	46%	
			25	**76%**	54%	60%	
3vs1	3	1	1	**51%**	38%	29%	3%
			5	**71%**	60%	46%	16%
			10	**79%**	68%	53%	30%
			25	**89%**	82%	66%	54%
5vs1	5	1	1	**64%**	52%	34%	
			5	**85%**	78%	61%	
			10	**93%**	86%	71%	
			25	**97%**	96%	87%	
3vs3	3	3	1	**70%**	62%	51%	12%
			5	**91%**	83%	80%	36%
			10	**95%**	90%	88%	57%
			25	**99%**	96%	96%	91%

Table 4 Comparison with the state of the art. Average accuracy at ranks 1, 5, 10 and 25 using RGB histograms with Mixture for both single shot and multi-shot cases

of the approach. Data source calibration mitigates the problem, although it is not required. Finally, since re-identification relies on the correspondence of color appearance, the system is strongly sensitive to differences in color responses of the cameras. In this case, color calibration is mandatory.

Results both in real surveillance videos and in a proposed benchmark dataset called 3DPeS are very promising. In this dataset, standard approaches based on 2D models fail, since the points of view are very different and automatic segmentation is not sufficiently accurate. We believe that this new way explored based on 3D body models could be the starting point for future innovative solutions. Spatial mapping of appearance features using a fixed or, even better – an articulated body model – detection and handling of discriminative details and the introduction of learning steps for dimensionality reduction or metric learning are the most promising research directions.

References

1. Vezzani, R., Baltieri, D., Cucchiara, R.: People re-identification in surveillance and forensics: a survey. ACM Computing Surveys **46** (2014)
2. Gong, S., Cristani, M., Yan, S., Loy, C., eds.: Person Re-Identification. Volume XVIII of Advances in Computer Vision and Pattern Recognition. Springer London (2014)
3. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: Proc. of CVPR. (2010) 623 –630
4. Colombo, C., Del Bimbo, A., Valli, A.: A real-time full body tracking and humanoid animation system. Parallel Comput. **34** (2008) 718–726
5. Doretto, G., Sebastian, T., Tu, P.H., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. J. Ambient Intelligence and Humanized Computing **2** (2011) 127–151
6. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S.: Multi-camera multi-person tracking for EasyLiving. In: Proc. Third IEEE Int. Workshop on Visual Surveillance, IEEE Comput. Soc (2000) 3–10
7. Cai, Q., Aggarwal, J.: Tracking human motion in structured environments using a distributed-camera system. IEEE Transactions. on Pattern Analysis and Machine Intelligence **21** (1999) 1241–1247
8. Albu, A., Laurendeau, D., Comtois, S., Ouellet, D., Hebert, P., Zaccarin, A., Parizeau, M., Bergevin, R., Maldague, X., Drouin, R., Drouin, S., Martel-Brisson, N., Jean, F., Torresan, H., Gagnon, L., Laliberte, F.: MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras. In: Proc. of ICPR, IEEE (2006) 924–928
9. Schügerl, P., Sorschag, R., Bailer, W., Thallinger, G.: Object re-detection using SIFT and MPEG-7 color descriptors. Lecture Notes In Computer Science (2007) 305–314
10. Bird, N., Masoud, O., Papanikolopoulos, N., Isaacs, A.: Detection of Loitering Individuals in Public Transportation Areas. IEEE Transactions. on Intelligent Transportation Systems **6** (2005) 167–177
11. Monari, E., Maerker, J., Kroschel, K.: A Robust and Efficient Approach for Human Tracking in Multi-camera Systems. In: Proc. of AVSS, IEEE (2009) 134–139
12. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proc. of CVPR. (2010) 2360–2367
13. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using spatial covariance regions of human body parts. In: Proc. of AVSS. (2010) 435–440
14. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: Proceedings of the 10th Asian conference on Computer vision - Volume Part IV. ACCV'10, Berlin, Heidelberg, Springer-Verlag (2011) 501–512
15. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: Proc. of CVPR. (2011) 649–656
16. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification.

In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: Computer Vision – ECCV 2012. Volume 7577 of Lecture Notes in Computer Science., Springer (2012) 780–793
17. Gandhi, T., Trivedi, M.: Panoramic Appearance Map (PAM) for Multi-camera Based Person Re-identification. In: Proc. of AVSS, IEEE (2006) 78–78
18. Barbosa, I.B., Cristani, M., Bue, A.D., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In Fusiello, A., Murino, V., Cucchiara, R., eds.: ECCV Workshops (1). Volume 7583 of Lecture Notes in Computer Science., Springer (2012) 433–442
19. Baltieri, D., Vezzani, R., Cucchiara, R.: Sarc3d: a new 3d body model for people tracking and re-identification. In: Proc. of IEEE Int. Conf. on Image Analysis and Processing, Ravenna, Italy (2011) 197–206
20. Baltieri, D., Vezzani, R., Cucchiara, R.: People orientation recognition by mixtures of wrapped distributions on random trees. In: Proceedings of the 12th European Conference on Computer Vision, Firenze, Italy (2012)
21. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Proceedings of the 1st International ACM Workshop on Multimedia access to 3D Human Objects, Scottsdale, Arizona, USA (2011) 59–64
22. Roullot, E.: A unifying framework for color image calibration. Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on (2008) 97–100
23. Gijsenij, A., Gevers, T., van de Weijer, J.: Computational color constancy: Survey and experiments. IEEE Transactions on Image Processing **20** (2011) 2475 –2489
24. Porikli, F.: Inter-camera color calibration by correlation model function. In: Proc. of ICIP. Volume 2. (2003) II – 133–6 vol.3
25. Javed, O., Shafique, K.: Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras. In: Proc. of CVPR, IEEE (2005) 26–33
26. Gilbert, A., Bowden, R.: Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In: Proc. of ECCV. (2006) 125–136
27. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of CVPR. Volume 1., Washington, DC, USA, IEEE Computer Society (2005) 886–893
28. Gualdi, G., Prati, A., Cucchiara, R.: Multi-stage sampling with boosting cascades for pedestrian detection in images and videos. In: Proc. of ECCV, Crete, Greece (2010) 196–209
29. Vezzani, R., Grana, C., Cucchiara, R.: Probabilistic people tracking with appearance models and occlusion classification: The ad-hoc system. Pattern Recognition Letters **32** (2011) 867–877
30. Baltieri, D., Vezzani, R., Cucchiara, R.: 3d body model construction and matching for real time people re-identification. In: Proceedings of Eurographics Italian Chapter Conference 2010 (EG-IT 2010), Genova, Italy (2010)
31. Vezzani, R., Cucchiara, R.: Video surveillance online repository (visor): an integrated framework. Multimedia Tools and Applications **50** (2010) 359–380
32. Kotoulas, L., Andreadis, I.: Colour histogram content-based image retrieval and hardware implementation. IEE Proceedings on Circuits, Devices and Systems, **150** (2003) 387–93–
33. Grana, C., Vezzani, R., Cucchiara, R.: Enhancing hsv histograms with achromatic points detection for video retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR 2007. (2007) 302–308
34. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press, ISBN: 0521540518 (2004)
35. Baltieri, D., Vezzani, R., Cucchiara, R.: Sarc3d: a new 3d body model for people tracking and re-identification. In: Proc. of IEEE Int. Conf. on Image Analysis and Processing, Ravenna, Italy (2011) 197–206
36. Jungling, K., Arens, M.: View-invariant person re-identification with an implicit shape model. In: Proc. of AVSS. (2011) 197 –202
37. Calderara, S., Cucchiara, R., Prati, A.: Bayesian-competitive consistent labeling for people surveillance. IEEE Transactions. on Pattern Analysis and Machine Intelligence **30** (2008) 354–360
38. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63** (2006) 3–42
39. Enzweiler, M., Gavrila, D.M.: Integrated pedestrian classification and orientation estimation. In: CVPR. (2010) 982–989
40. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR. (2010) 623 –630
41. Hellinger, E.: Neue begrndung der theorie quadratischer formen von unendlichvielen vernderlichen. J. Reine Angew. Math. **136** (1909) 210–271
42. Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In: Proc. of 10th IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS). (2007)
43. Schwartz, W., Davis, L.: Learning Discriminative Appearance-Based Models Using Partial Least Squares. In: Proc. of the XXII Brazilian Symposium on Computer Graphics and Image Processing. (2009)
44. Vezzani, R., Cucchiara, R. In: Benchmarking for Person Re-identification. Springer London (2014) 333–349
45. Nilski, A.: Evaluating multiple camera tracking systems - the i-lids 5th scenario. In: Security Technology, 2008. ICCST 2008. 42nd Annual IEEE International Carnahan Conference on. (2008) 277 –279
46. Bk, S., Corve, E., Brmond, F., Thonnat, M.: Boosted human re-identification using riemannian manifolds. Image and Vision Computing **30** (2012) 443 – 452
47. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: Proc. of ECCV, Berlin, Heidelberg, Springer-Verlag (2008) 262–275
48. Gandhi, T., Trivedi, M.M.: Person tracking and reidentification: Introducing Panoramic Appearance Map (PAM) for feature representation. Machine Vision and Applications **18** (2007) 207–220