SARC3D: a new 3D body model for People Tracking and Re-identification

Davide Baltieri, Roberto Vezzani, and Rita Cucchiara

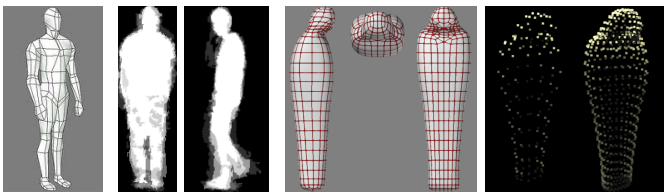
Dipartimento di Ingegneria dell'Informazione - University of Modena and Reggio Emilia, Via Vignolese, 905 - 41125 Modena - Italy
{davide.baltieri,roberto.vezzani,rita.cucchiara}@unimore.it

Abstract. We propose a new simplified 3D body model (called Sarc3D) for surveillance application, that can be created, updated and compared in rea-time. People are detected and tracked in each calibrated camera, and their silhouette, appearance, position and orientation are extracted and used to place, scale and orientate a 3D body model. Foreach vertex of the model a signature (color features, reliability and saliency) is computed from the 2D appearance images and exploited for mathing. This approach achieves robustness against partial occlusions, pose and viewpoint changes. The complete proposal and a full experimental evaluation is presented, using a new benchmark suite and the PETS2009 dataset.

Key words: 3D human model, People Re-identification

1 Introduction and related work

People Re-identification is a fundamental task for the analysis of long-term activities and behaviors of specific people. Algorithms have to be robust in challenging situations, like widely varying camera viewpoints and orientations, varying poses, rapid changes in part of clothes appearance, occlusions, and varying lighting conditions. Moreover people re-identification requires elaborate methods in order to cope with the widely varying degrees of freedom of a person's appearance. Various algorithms have been proposed in the past, based on the kind of data available: a first category of person re-identification methods relies on biometric techniques, such as face [1] or gait, but high resolution or PTZ cameras are required in this case. Other approaches suppose well constrained operative conditions, calibrated cameras and precise knowledge of the geometry of the scene; the problem is simplified by adding spatial and/or temporal constraints in order to greatly reduce the candidate set [2–4]. Finally, most re-identification methods purely rely on appearance-based features; a comparison and evaluation of some of them is reported in [5, 6]. For example, Farenzena et el [6] proposed to divide the person appearance into five parts using a rule based approaches to detect head, torso and legs and image symmetries to split torso and leg regions into left and right ones. For each region, a set of color and texture features are collected for the matching step. Alahi *et al* [7] proposed a general framework for simultaneous tracking and re-detection by means of a grid cascade of dense



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

Fig. 1. (a) a human 3d model, (b) average silhouettes used for the model creation, (c) our simplified human model, (d) the vertices sampling used in our tests

region descriptors. Various descriptors have been evaluated, like SIFT, SURF and covariance matrices, and the latter are shown to outperform the formers. Finally, [8] proposed the concept of Panoramic Appearance Map to perform re-identification. This map is a compact signature of the appearance information of a person extracted from multiple cameras, and can be thought of as the projection of a person appearance on the surface of a cylinder.

In this paper we present the complete design of a method for people re-identification based on a 3D body models. The adoption of 3D body models is new for re-identification, differently from other computer vision fields, such as motion capture and posture estimation [9, 10]. The challenges connected with 3D models rely on the need of precise people detection, segmentation and estimation of the 3D orientation for a correct model to image alignment. However, our proposal has several benefits; first of all, we provide an approximate 3D body model with a single shape parameter, which can be learned and used for comparing people from very few images (even only one); due to the precise 3D feature mapping, the comparison allows to look for details, and not only to global features, it also allows to cope with partial data and occluded views; finally, the main advantage of a 3D approach is to be intrinsically independent of the point of view. Our approximate body model (called Sarc3D) is not only fast but also suitable for low resolution images as the ones typically acquired by surveillance cameras. A preliminary version of this work was presented in [11], while in this paper we present the complete proposal together with a comprehensive set of experiments, both on a new benchmark suite, available with the corresponding annotation, and on real scenarios using the PETS2009 dataset.

2 The Sarc3D Model

First of all the generic body model of a person must be defined. Differently from motion capture or action recognition systems, we are not interested in the precise location and pose of each body part, but we need to correctly map and store the person appearance. Instead of an articulated body model (as in fig. 1(a)), we propose a new monolithic 3d model, called "Sarc3D". The model construction has been driven by real data: side, frontal and top views of generic

people were extracted from various surveillance videos; thus, for each view an average silhouette has been computed and used for the creation of a graphical 3d body model (see fig. 1(b)), producing a sarcophagus-like (fig. 1(c)) body hull. The final body model is a set of vertices regularly sampled from the sarcophagus surface. The number of sampled vertices could be selected accordingly to the required resolution. In our tests on real surveillance setups, we used from 153 to 628 vertices (fig. 1(d)). Other sampling densities of the same surface have been tested, but the selected ones outperformed the others on specificity and precision tests, and are a good trade-off between speed and efficacy. As a representative signature, we created an instance Γ^p of the generic model for each detected person p-th, characterized by a scale factor (to cope with different body builds) and relating appearance information (i.e., color and texture) to each vertex, defined as:

$$\Gamma^p = \{h^p, \{v_i^p\}\}, p \in [1 \dots P], i \in [1 \dots M] \quad (1)$$

where h^p is the person height, as extracted by the tracking module, and used as the scale factor for the 3D model; v_i^p is the vertex set; P is the number of people in the gallery and M is the number of vertices. For each vertex the following five features are computed and stored:

- $\boldsymbol{n}_i$: the normal vector of the 3D surface computed at the i-th vertex location; this feature is pre-computed during the sampling of the model surface;
- c_i: the mean color;
- $\mathbf{H_i}$: a local HSV histogram describing the color appearance of the vertex neighbor; it is a normalized histogram with 8 bins for the *hue* channel and 4 bins for the *saturation* and *value* channels respectively;
- θ_i: the optical reliability value of the vertex, which takes into account how well and precisely the vertex color and histogram have been captured from the data.
- s_i: the saliency of the vertex, which indicates its uniqueness with respect to the other models; i.e., the saliency of a vertex will be higher in correspondence to a distinctive logo and lower on a common jeans patch.

2.1 Positioning and orientation

The 3D placement of the model in the real scene is obtained from the output of a common 2D surveillance system working on a set of calibrated cameras. Assuming a vertical standing position of the person, the challenging problem to solve is the estimation of his horizontal orientation. To this aim, we consider that people move forward and thus we exploit the trajectory on the ground plane to give a first approximation using a sliding window approach. Given a detected person, we consider a window of N frames and the corresponding trajectory on the ground plane. A quadratic curve is then fitted on the trajectory and the fit score is used as orientation reliability. If it is above a predefined threshold, the final orientation is generated from the curve tangent.

A finer angle adjustment is provided by a generative approach using the already computed part of the 3D model (if available). In fig.2(a) and 2(b) a sample

frame and the corresponding model placement and orientation is provided. In particular, the sample positions used for the curve fitting and orientation estimation are highlighted.

2.2 Model creation

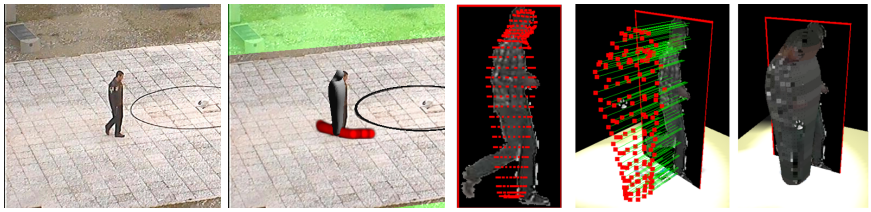
Given the 3D placement and orientation, the appearance part of the model can be recovered from the 2D frames. Projecting each vertex v_i to the camera image plane, the corresponding nearest pixel $x(v_i), y(v_i)$ is obtained. The vertex color c_i is initialized directly using the image pixel:

$$c_i = I\left(x(v_i), y(v_i)\right); \tag{2}$$

where I is the analyzed frame, $x(v_i)$ and $y(v_i)$ are the image coordinates of the projection of the vertex v_i. The histogram $\mathbf{H_i}$ is computed on a squared image patch centered around $(x(v_i), y(v_i))$. The size of the patch is selected taking into account the sampling density of the 3D model surface and the mean size of the blobs items. In our experiments, we used 10x10 blocks. Finally, the optical reliability value is initialized as $\theta_i = \boldsymbol{n}_i \cdot \boldsymbol{p}$, where $\boldsymbol{p}$ is the normal to the image plane (equal to the inverted direction vector of the camera). The reason behind the adoption of the dot product is that data from front-viewed vertices and their surrounding surface are more reliable than from lateral viewed vertices. The vertices belonging to the occluded side of the person are also projected onto the image, but their reliability has a negative value due to the opposite directions of $\boldsymbol{n}_i$ and $\boldsymbol{p}$. Thus each vertex of the model is initialized even with a single image: from a real view if available or using a sort of symmetry-based hypothesis in absence of information. However, negative values of the reliability allow to identify vertices initialized with a forecast and not directly from the data. The vertices having no match with the current image (i.e. the vertices projected outside of the person silhouette) are also initialized with a copy of the feature vector of the nearest initialized vertex and their reliability values are set to zero. By means of the reliability value vertices directly seen at least once ($\theta > 0$), vertices initialized using a mirroring hypothesis ($-1 \leq \theta < 0$) and vertices initialized from its neighborhood ($\theta = 0$) are distinguishable. The described steps of the initialization phase are depicted in Fig. 2(c).

If multiple cameras are available or if the short-term tracking system provides more detections for the same object, the 3D model could integrate all the available frames. For each of them, after the alignment step, a new feature vector is computed for each vertex successfully projected inside the silhouette of the person. The previously stored feature vector is then merged or overwritten with the new one, depending on the signs of the reliabilities. In particular, direct measures ($\theta > 0$) always overwrite forecasts ($\theta < 0$), otherwise they are merged as in the following equations:

$$\hat{c}_i^p = \frac{\theta_i^p c_i^p + \theta_i^s c_i^s}{\theta_i^p + \theta_i^s}, \qquad \hat{\mathbf{H}}_i^p = \frac{\theta_i^p \mathbf{H_i^P} + \theta_\mathbf{i}^\mathbf{s} \mathbf{H_i^s}}{\theta_i^p + \theta_i^s}, \qquad \hat{\theta}_i^p = \frac{\theta_i^p + \theta_i^s}{2} \tag{3}$$



(a) (b) (c)

Fig. 2. (a) A frame from a video, (b) Automatic 3D positioning and orientation (c) Initialization of the 3D model of a person: the model to image alignment, projection of the model vertex to the image plane, vertex initialization or update

The normal vector $\boldsymbol{n}_i^p$ does not change in the merging operation, since it is constantly obtained during the model generation, while the saliency s_i^p is recomputed after the merging. Figure 3 shows some sample models created from one or more images.

2.3 Occlusion management and view selection

Not all views should be used for the initialization and update of the model. Errors in the tracking step, noise and bad calibration could lead the model to become degraded. To this aim a rule based approach is proposed to select and exploit the best views only for the model initialization and update.

Occlusion check: In addition to 2D occlusion detection algorithms [12], a computer graphic based generative approach is used: for the selected camera view and for each person visible by that camera, a binary image mask $\hat{I}_p$ is rendered for each person p in the scene. An occlusion is detected each time two model masks are overlapping or connected. 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 Fig. 4.

Model to foreground overlapping: The reliability of the model positioning could be evaluated considering the overlapping area between the 2D foreground mask and the rendered images $\hat{I}_p$. For each person, the overlapping

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

Fig. 3. Various models created and the corresponding source images

(a) (b) (c)

Fig. 4. Occlusion detection: (a) the input frame, (b) the aligned 3D models and (c) the mask generated by the rendering system. Since the blue and green objects are connected, the corresponding models are frozen and not updated during the occlusion

score R_p is computed as the ratio between the number of foreground pixels that overlap with $\hat{I}_p$ with respect to the total number of silhouette pixels. If R_p is higher than a strong threshold (e.g., 95% in out experiments) the selected view is marked as good. Otherwise the alignment is not precise enough or the person is not assuming a standing position compliant with the sarcophagus model.

Orientation reliability: As before-mentioned, the reliability of the orientation estimation could be evaluated considering the fitting score of the quadratic curve as described in Section 2.1. A similar check could be performed 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 not reliable.

If all above conditions hold true, the estimated orientation and position are considered reliable and the selected view could be exploited to initialize (or update) the model.

3 People re-identification

One of the main applications of the Sarc3D model is the people re-identification. Goal of the task is to find possible matches among couples of models from a given set of SARC3D items. First, we define the distance between two feature vectors. Using the optical reliability θ_i and the saliency s_i as weighting parameters, the Hellinger distance between histograms is used:

$$d(v_i^p, v_i^q) = d_{He}\left(\mathbf{H_i^p}, \mathbf{H_i^q}\right) = \sqrt{1 - \sum_{h,s,v} \sqrt{H_i^p(h,s,v) \cdot H_i^q(h,s,v)}}. \tag{4}$$

The distance $D_H(\Gamma^p, \Gamma^q)$ between two models Γ^p and Γ^q is the weighted average of the vertex-wise distances, using the product of the reliabilities as weight.

$$D_H(\Gamma^p, \Gamma^t) = \frac{\sum_{i=1\ldots M}\left(w_i \cdot d(v_i^p, v_i^q)\right)}{\sum_{i=1\ldots M}\left(w_i\right)} \tag{5}$$

where

$$w_i = f(\theta_i^p) \cdot f(\theta_i^q) \tag{6}$$

Thus generic global distance assumes that each vertex has the same importance and the weights w_i are based only on optical properties of the projections or the reliability of the data. We believe that global features are useful to reduce the number of candidates or if the resolution is low. However, the final decision should be guided by original patterns and details, as humans normally do to recognize people without biometric information (e.g., a logo in a specific position of the shirt). To this aim we have enriched the vertex feature vector v_i^p with a saliency measure $s_i^p \in [0 \dots 1]$. Given a set of body models, the saliency of each vertex is related to its minimum distance from all the corresponding vertices belonging to the other models:

$$s_i^p \propto \min_t \left(d_H(\mathbf{H_i^p}, \mathbf{H_i^t}) \right) + s_0, \qquad \sum s_i^p = 1 \tag{7}$$

where s_0 is a fixed parameter to give a minimum saliency to each vertex. If s is low, the vertex appearance is not distinctive; otherwise, the vertex has completely original properties and it could be used as a specific identifier of the person. The corresponding saliency-based distance D_S can be formulated based on new weights by substituting w_i' to w_i in eq.6.

$$w_i' = f(\theta_i^p) \cdot f(\theta_i^q) \cdot s_i^p \cdot s_i^q \tag{8}$$

This saliency-based distance D_S cannot be used instead of Eq. 5, since it focuses on details discarding global information and then leading to macroscopic errors; the re-identification should be done based on both global (D_H) and local (D_S) similarities. Thus, the final distance measure D_{HS} used for re-identification is the product of the two contributions $D_{HS} = D_H \cdot D_S$.

4 Experimental results

Many experiments have been carried out, on real videos and on our benchmarking suite. From its introduction, ViPER [5] is the reference dataset for re-identification problems. Unfortunately, it contains two images for each target only. Thus we propose a suitable benchmark dataset[1] with 50 people, consisting of short video clips captured with a calibrated camera. The annotated data set is composed by four views for each person, 200 snapshots in total. Some examples are shown in Fig. 5(a), where the output of the foreground segmentation is reported.

For each testing item we ranked the training gallery using the distance metrics defined in the previous section. The summarized performance results are reported using the cumulative matching characteristic (CMC) curve [5]. Each test was replicated exhaustively choosing different combinations of images.

In fig.6 we report the performance obtained using the proposed distances on Sarc3D models. Three different images for each person were chosen as training set (i.e. for the model creation) while the remaining form the test set. Each test

[1] available here: `http:\\imagelab.ing.unimore.it\sarc3d`

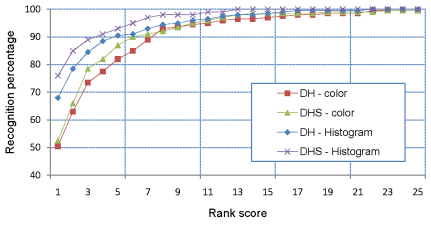
			
	**0.795246**	0.659210	0.566881
	0.639717	**0.672695**	0.656614
	0.510284	0.530336	**0.727577**

(a) (b)

Fig. 5. (a) Selected images from the benchmarking suite, (b) distance matrix obtained considering three very similar people: the three images used for the model creation (rows) and the test images (column) are also shown.

was replicated four times using different split of the images into training and testing sets. For each testing image, the ranking score was obtained considering the model-to-model matching schema using the histogram based distance D_H and the saliency-based one D_{HS}. For sake of comparison, we have evaluated the results obtained using the Euclidean distance between image pixels and the vertex mean color c_i instead of the histogram-based one of Eq. 4. In Table **??** some key values extracted from the graphs are reported, showing the performance improvements obtained using 3D-3D model matching instead of 3D-2Dm easures and, in the last row, adopting the saliency measure.

In this paper we assumed to have a sufficiently accurate tracking system, which gives the 2D foreground images used for the model alignment. However, the proposed method is reliable and robust enough, even in case of approximated alignments. The use of a generic sarcophagus-like model and local color histograms instead of detailed 3D models and point-wise colors goes precisely in this direction and allow to cope with small alignment errors. In addition, the introduction of the normal vector to the 3D surface assign strong weights to the central points of the people and very low weights to lateral points, which are the most hit by misalignments.



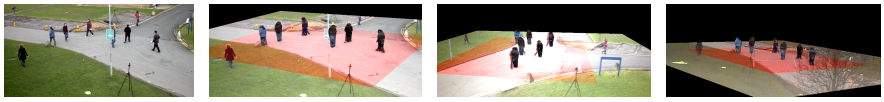
Test	Rank-1	Rank-2	Rank-5
3D-2D (1 image, D_H)	50.0%	70.0%	85.0%
3D-2D (2 images, D_H)	60.3%	73.2%	86.6%
3D-2D (3 images, D_H)	68.0%	78.5%	90.5%
3D-3D (2 images, D_H)	74.0%	82.0%	91.0%
3D-2D (3 images, D_{HS})	76.0%	85.0%	93.0%

(a) (b)

Fig. 6. Experimental results of the 3 image model-to-image test with the D_{HS} distance

	Correct	With noise on localization					With noise on orientation						
	Alignment	2 px	4 px	6 px	12 px	16 px	5°	10°	15°	30°	40°	60°	90°
Rank		(15%)	(30%)	(45%)	(75%)	(90%)							
1	0.68	0.63	0.60	0.58	0.51	0.45	0.61	0.60	0.58	0.55	0.54	0.53	0.50
2	0.78	0.75	0.73	0.72	0.65	0.60	0.74	0.75	0.76	0.70	0.69	0.69	0.66
3	0.83	0.83	0.81	0.80	0.74	0.68	0.80	0.81	0.81	0.77	0.75	0.81	0.75
4	0.85	0.86	0.85	0.84	0.76	0.75	0.85	0.84	0.83	0.81	0.79	0.85	0.80
5	0.90	0.88	0.86	0.85	0.78	0.78	0.89	0.87	0.88	0.86	0.82	0.87	0.84
10	0.96	0.96	0.97	0.95	0.94	0.90	0.96	0.96	0.96	0.94	0.95	0.96	0.95

Table 1. Performance evaluation of the system using random perturbation of the 3D model localization and orientation (3D-2D matching)



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

Fig. 7. (a) PETS dataset, sample frame from camera 1, (b,c,d) system output superimposed to camera 1, 2 and 3 frames

In table 1 the performance of the system in presence of random perturbations of the correct alignment is reported; both errors on the ground plane localization and on the orientation have been introduced. The performance reported on the table shows that our system is still reliable, even in the case of non precise model alignment and orientation, keeping good results with localization precision up to 6 pixels (45% overlap between the projected bounding box of the model and the image blob) and orientation up to 30 degrees.

We also tested the system on the PETS 2009 dataset [13] to evaluate the proposed method in real life conditions. The *City center* sequence, with three overlapping camera views, was selected. A 12.2m $\times$ 14.9m ROI was choosed, which is visible from all cameras. The proposed method was added on top of a previously developed tracking system [14], the goal of our method was to repair broken track and re-identify people that enter and exit the rectangular ROI. The D_H distance is used, together with the 153-vertices model, since the low resolution. Fig. 7 shows some sample frames from the system in action. The obtained precision and recall are 80.2% and 88.7

5 Conclusions

We proposed a new and effective method for people re-identification. Differently from currently available solutions we exploited a 3D body model to spatially localize identifying patterns and colors on the vertices of the model. In this way, occlusion and view dependencies are intrinsically solved. Results both in real surveillance videos and in a proposed benchmark dataset are very promising. In

this dataset, standard approaches based on 2D models fail, since the points of view are very different and the automatic segmentation is not precise enough. We believe that this new explored way based on 3D body models could be the starting point for future innovative solutions.

Acknowledgments

The work is currently under development as a doctorate project of the ICT School of the University of Modena and Reggio Emilia, and within the EU project THIS (DG - JLS).

References

1. Bäuml, M., Bernardin, K., Fischer, M., Ekenel, H., Stiefelhagen, R.: Multi-Pose Face Recognition for Person Retrieval in Camera Networks. In: Proc. of AVSS. (2010)
2. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding **109** (2008) 146–162
3. Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: Proc. of CVPR. (2004) 205–210
4. Vezzani, R., Baltieri, D., Cucchiara, R.: Pathnodes integration of standalone particle filters for people tracking on distributed surveillance systems. In: Proc. of ICIAP, Vietri sul Mare, Italy (2009)
5. Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In: Proc. of PETS 2007. (2007)
6. 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
7. Alahi, A., Vandergheynst, P., Bierlaire, M., Kunt, M.: Cascade of descriptors to detect and track objects across any network of cameras. Computer Vision and Image Understanding **114** (2010) 624–640
8. Gandhi, T., Trivedi, M.: Panoramic Appearance Map (PAM) for Multi-camera Based Person Re-identification. In: Proc. of AVSS. (2006) 78–78
9. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: Proc. of CVPR. (2010) 623 –630
10. Colombo, C., Del Bimbo, A., Valli, A.: A real-time full body tracking and humanoid animation system. Parallel Comput. **34** (2008) 718–726
11. Baltieri, D., Vezzani, R., Cucchiara, R.: 3D body model construction and matching for real time people re-identification. In: Proc. of Eurographics Italian Chapter Conference (EG-IT 2010), Genova, Italy (2010)
12. 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
13. PETS: Dataset - Performance Evaluation of Tracking and Surveillance (2009)
14. Vezzani, R., Cucchiara, R.: Event driven software architecture for multi-camera and distributed surveillance research systems. In: Proceedings of the First IEEE Workshop on Camera Networks - CVPRW, San Francisco (2010) 1–8