Learning Articulated Body Models for People Re-identification

Davide Baltieri, Roberto Vezzani, Rita Cucchiara
University of Modena and Reggio Emilia
Via Vignolese 905, 41125 Modena - Italy
{davide.baltieri, roberto.vezzani, rita.cucchiara}@unimore.it

ABSTRACT

People re-identification is a challenging problem in surveillance and forensics and it aims at associating multiple instances of the same person which have been acquired from different points of view and after a temporal gap. Image-based appearance features are usually adopted but, in addition to their intrinsically low discriminability, they are subject to perspective and view-point issues. We propose to completely change the approach by mapping local descriptors extracted from RGB-D sensors on a 3D body model for creating a view-independent signature. An original bone-wise color descriptor is generated and reduced with PCA to compute the person signature. The virtual bone set used to map appearance features is learned using a recursive splitting approach. Finally, people matching for re-identification is performed using the Relaxed Pairwise Metric Learning, which simultaneously provides feature reduction and weighting. Experiments on a specific dataset created with the Microsoft Kinect sensor and the OpenNi libraries prove the advantages of the proposed technique with respect to state of the art methods based on 2D or non-articulated 3D body models.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Computer Vision*

General Terms

Algorithms,Experimentation

Keywords

People re-identification, Microsoft Kinect, RGB-D sensors

1. INTRODUCTION

People re-identification is a challenging task extremely useful for video surveillance or forensics. It aims at discovering multiple instances of the same person captured from different points of view or after a significant temporal gap. However, differently from biometric techniques, re-identification trusts on appearance information only and thus it is mainly based on the color, texture and shape of people clothing [4]. Features such as color and texture histograms have been deeply tested and applied rather than shape and geometrical properties, leading to an intrinsically low discriminability of the computed signatures.

To solve this problem, 2D models and, recently, non-articulated 3D models of the human body have been introduced. Differently from motion capture or action analysis approaches, the models are not required to be extremely precise but fast. 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 segmentation errors can also be minimized.

In this work we made a step forward by proposing a new original 3D approach to re-identification based on articulated body models. A 3D model is adopted to map appearance descriptors to skeleton bones (See Fig. 1). The color, depth and skeleton streams produced with the Microsoft Kinect sensor [11] and the OpenNi libraries are exploited as input. The skeleton is further refined using a learning approach in order to generate a "bone" set. The obtained signature is strictly related to the real body structure, thus it also allows an attribute based description, which could be useful is some applications [8]. In addition, a learned metric is adopted which simultaneously acts as feature selection and body part weighting.

2. RELATED WORKS

Simplified *2D models* are the most commonly used in people re-identification. Among the others, in surveillance and forensics the *cylindrical* shape and the *legs-torso-head* structure are the most widely adopted. By modeling a person as a cylindrical shape (or more generally as a solid of revolution) the horizontal variations of the people appearance are neglected, supposing that the color or texture distribution along the vertical axis is the only important

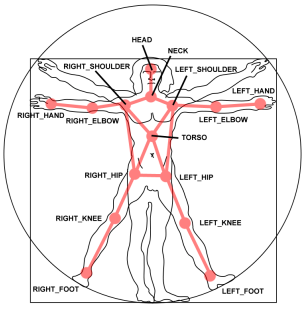


Figure 1: Vitruvian body model with superimposed the joints and bones used in the proposed system

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502147.

data. For example, in [3] the person mask is divided into ten horizontal stripes (ideally corresponding to legs, torso, and head) and the mean color of each stripe is stored as representative feature. Weber *et al* adopted a part based detector to separately segment a personâĂŹs upper and lower clothing regions [13]. Similarly, Farenzena *et al* [5] compute the cut points from profile histograms. Moreover, they operate a further split in two of the torso and legs parts using a symmetry based algorithm.

Very recently, simplified *3D models* have been proposed to improve the feature localization. For example, a non-articulated 3D model called Sarc3D [1] has been proposed to map color appearance features on a vertex based model. The main drawback of Sarc3D is the lack of articulation. The rigid container requires a corresponding rigid posture of the person, while it will fail on people in sitting or bending positions. Articulated motion capture has been addressed in the past on color images [10, 9] and, more recently, using depth images [11, 12], but never applied to re-identification. Recently, the introduction of low cost range sensors like the Microsoft Kinect opened new solutions to surveillance and forensics research fields, included the re-identification task. Soft-biometry features can be extracted from the tracked skeleton stream and used to generate a person profile. For example, in [2] a set of ratios of joint distances is used as person signature. However, the intrinsic noise on the estimation of the joint positions do not allow to reach very high performance.

3. BONE FEATURE SET AND SIGNATURE

In order to provide a spatial location to each appearance feature, pixels of the color image need to be connected to a bone of the person. To this aim, OpenNi is exploited to find a set of human joints (see the red dots in Figure 1). Then these joints are linked together to build a simple human skeleton (see Figure 1). Let be $\Psi = \{\psi_1, \ldots, \psi_{15}\}$ the set of 15 joints, where each of the elements ψ_i corresponds to the 3D position of the joint. Based on Ψ, the corresponding "bone" set $\mathcal{B} \subset \Psi \times \Psi$ is obtained as the set of edges of the joint graph (see Figure 1). Each bone $b_i \in \mathcal{B}$ is defined using the 3D coordinates of the two extremities: $b_i = (\psi_r, \psi_s)$. Let $N = |\mathcal{B}|$ be the number of bones, 18 in our skeleton model.

Given the point cloud $\mathcal{W} = \{(x_1, c_1) \ldots (x_W, c_W)\}$, where x_j and c_j are the position and the color of each point respectively, a point-to-bone assignment is provided using a min-distance criteria and the subsets $\mathcal{W}_i$ of points connect to the i-th bone b_i are obtained as follows:

$$\mathcal{W}_i = \{(x_j, c_j) \in \mathcal{W} | i = \phi(x_j)\}, j = 1 \ldots W \tag{1}$$

where ϕ is the function returning the index of the closest bone:

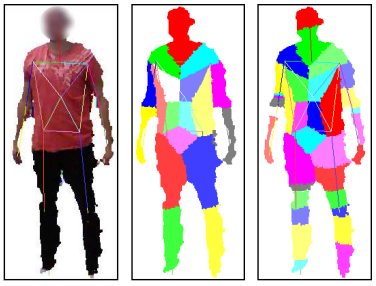
$$\phi(x_j) = \arg\min_{i=1\ldots N} d(x_j, b_i) \tag{2}$$

The pixel-to-bone distance $d(x_j, b_i)$ is the common point-to-segment Euclidean distance.

After the pixel-to-bone assignment, the signature of the person is composed by the set of color histograms computed for each bone:

$$\mathbf{H}^p = \{H_1^p, \ldots, H_N^p\}. \tag{3}$$

where H_i^p is the color histogram for the bone b_i of the p-th person. In the experiments reported in the paper, H_i^p are RGB color histograms with a 8 bin quantization for each channel, normalized to sum up to 1. If the person model is obtained as integration of multiple views, the histograms are computed using all the image points assigned to the same bone. A visual example of the pixel-to-bone assignment is reported in Figure 2(b).



(a) (b) (c)

Figure 2: Point cloud to bone assignment. a) the skeleton, b) the point assignment using the default bone set and c) after automatic fragmentation.

4. DISTANCE METRIC

The distance between two person signatures $\mathbf{H}^i$ and $\mathbf{H}^j$ can be computed as the sum of distances between each corresponding couple of histograms:

$$d(\mathbf{H}^i, \mathbf{H}^j) = \sum_{n=1}^{N} \alpha_n \cdot d(H_n^i, H_n^j), \tag{4}$$

where $d(H_n^i, H_n^j)$ is the Mahalanobis distance between H_n^i and H_n^j. However, the distance function of Eq. 4 requires the estimation of the bones weights α_n. Moreover, the dimensionality of the signature $\mathbf{H}^p$ computed for each person as in equation Eq. 3 is too high, leading to problems on its storage and matching. Using the defined skeleton model composed by 18 bones and using 512 bins for each histogram, the final signature is composed by 9216 elements. Each signature is thus processed with a PCA step, which simultaneously reduce the dimensionality and filter the intrinsic noise. The person signature obtained from $\mathbf{H}^p$ becomes a 96 dimensional feature vector $x_p = x(\mathbf{H}^p)$. The PCA subspace has been learned on the training set and it is kept fixed for all the experiments, in order to obtain comparable signatures.

Instead of using Eq. 4, the metric scheme we adopted to compare person signatures is inspired from the one presented in [7], called Relaxed Pairwise Metric Learning (RPML), that has proved to be a highly efficient and effective metric learning approach. RPML aims at computing a pseudo-metric M, which, similarly to the Mahalanobis distance, provides a dissimilarity score of two feature vectors x_i and x_j:

$$\begin{aligned} d_M(\mathbf{H}^i, \mathbf{H}^j) &\doteq d_M(x_i, x_j) \\ &= (x_i - x_j)^\top M (x_i - x_j) = \\ &= \| L(x_i - x_j) \|^2 \end{aligned} \tag{5}$$

where $M = L^\top L$.

In order to exploit the discriminative information of the data during the metric learning, the person re-identification problem is redefined as a two-class problem: firstly, samples from the data space are converted to the label agnostic difference space. Secondly, the original class labels are discarded and the sample are rearranged into the *similar* and *different* classes $\mathcal{S}$ and $\mathcal{D}$ (i.e., if two signatures belong to the same person, their difference is labeled as $\mathcal{S}$, otherwise as $\mathcal{D}$).

Starting from the following objective function:

$$\mathcal{L}(L) = \frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}||L(x_i - x_j)||^2 - \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}}||L(x_i - x_j)||^2 \quad (6)$$

the problem of finding the best M can then be defined as

$$\begin{aligned} \arg\min \quad & \mathcal{L}(M) \\ \text{subject to} \quad & M \succeq 0, \\ & L\Sigma_S L^\top = L\Sigma_D L^\top = I \end{aligned} \quad (7)$$

where

$$\mathcal{L}(M) = tr(M(\Sigma_S - \Sigma_D)) \quad (8)$$

and

$$\Sigma_S = \frac{1}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}(x_i - x_j)(x_i - x_j)^\top \quad (9)$$

$$\Sigma_D = \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}}(x_i - x_j)(x_i - x_j)^\top \quad (10)$$

By relaxing the positivity constraint $M \succeq 0$, the problem can be simplified into finding M such that

$$tr(M(\Sigma_S - \Sigma_D)) = 0 \quad (11)$$

A bounded approximation for M ca be found as $M = ({\Sigma_S}^{-1} - {\Sigma_D}^{-1})$.

5. AUTOMATIC BONE FRAGMENTATION

The bone set $\mathcal{B}$ defined using the fifteen joints from OpenNi is not optimized for people re-identification. The main drawbacks are related to the different lengths of the bones and the semantic differences between bones and appearance regions. To this aim, an automatic bone fragmentation step has been introduced in order to generate a set of bones $\mathcal{B}^*$ more suitable for re-identification tasks. Our goal is to find the skeleton partitioning which maximizes the re-identification score (see Section 6). The size of the corresponding search space is too high for an exhaustive global maximum selection. An iterative procedure inspired to the Cuckoo search approach [14] has been implemented.

The set $\mathcal{B}^*$ is iteratively derived from $\mathcal{B}$ replacing one of the existing bones with the couple of sub-bones obtained after a split. A split $S = (b_i, \alpha)$ is represented by the index of the selected bone b_i and the split position $\alpha \in [0, 1]$. At each iteration, a set of candidates $\{S^k\}$ is randomly generated by selecting both the bone index and the split point α using a fast uniform distributed generator. The best split is then selected maximizing the re-identification score on a training set of samples. The iterative algorithm is stopped when the re-identification performance are not increased by adding an additional split. The final bone set created using our training set and the corresponding pixel-to-bone assignment are depicted in Figure 2(c) and is composed by 38 elements.

6. EXPERIMENTAL EVALUATION

In order to evaluate the proposed method we created a new dataset using Microsoft Kinect to extract depth information and relative point cloud. The dataset contains various images of 40 people in different poses and orientations, for a total of 450 shots and relative skeleton and point clouds. Half of the images were used for metric learning and as training set for the bone fragmentation learning, while the remaining 225 shots as testing. Some sample images from the dataset are reported in Figure 3.

Figure 3: Sample images and corresponding point clouds with the estimated skeletons from the Kinect dataset

As well-established in re-identification and recognition tasks, for each testing item we ranked the training gallery elements using the distance metric defined in the previous sections. The summarized performance results are reported using the the Cumulative Matching Characteristic (CMC) curve [6]. In order to evaluate the contribution of the bone fragmentation algorithm described in Section 5 and the metric learning defined in Section 4 we firstly performed an internal comparison. The results obtained using the base system (i.e., using the OpenNi bone set and the distance function of Eq. 4 with uniform weights) and with the integration of the two learning steps are reported in Figure 5(a). The curves have been obtained averaging 50 experiments with different training and the testing sets (randomly selected). As highlighted by the graph, the learned metric strongly improves the re-identification performance.

We also compared our method with two state of the art algorithms, namely SDALF [5] and Sarc3D [1]. SDALF is a purely two dimensional method. It consists in the extraction of features that model three complementary aspects of the 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 online code has been adopted and applied to the RGB images after the removal of the background by means of the depth stream. Sarc3D, instead, is a 3D method based on a non-articulated model. HSV histograms are extracted from patches of the foreground appearance image projected on a 3D surface. The tests have been carried out using our re-implementation of the method, exploiting the vertex model available online. The head and feet positions from the skeleton stream have been adopted as reference to align the model on the images. Results of the three methods are reported in Figure 5(b). The proposed method outperforms both the Sarc3D and the SDALF approaches. It requires an off-line training in order to learn the metric (which takes on average 4 seconds) and the best splits of the skeleton (195 minutes on average). The creation of a person model requires less than 10 ms (9.74ms for the computation of the point cloud by OpenNi and 0.47ms for the computation of the histograms), the matching score computation takes on average 2.11ms. All tests were performed on an Intel Core i5 running at 2.66 GHz. Some sample results of the three methods on the new dataset are shown in Figure 4, as can be seen our method is much more resilient to pose changes, unlike SDALF and Sarc3D which assume people in a vertical standing position.

7. CONCLUSIONS

We proposed a new and effective solution for the re-identification of people. Differently from currently available proposals, we exploited an articulated 3D body model to spatially localize the identifying patterns and colors on virtual bones. The articulated body model allows to create a invariant signature for each mon-

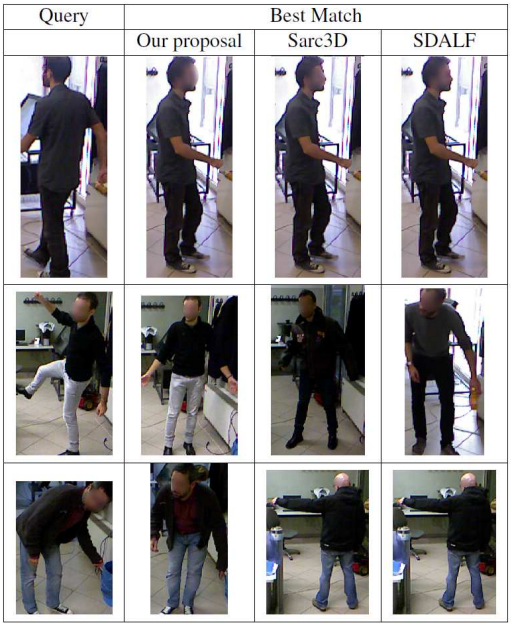


Figure 4: Sample results ot the three tested methods on our dataset

itored person despite postures and view-points. Two learning algorithms have been introduced to create an optimized bone set for re-identification purposes and a specific metric learning, respectively. Experimental results on a Microsoft Kinect dataset prove the improvement obtained using articulated models instead of fixed containers (e.g., Sarc3D [1]) or 2D body models (e.g., SDALF [5]).

8. REFERENCES

[1] D. Baltieri, R. Vezzani, and R. Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *Proc. of IEEE Int. Conf. on Image Analysis and Processing*, pages 197–206, Ravenna, Italy, Sept. 2011.

[2] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *ECCV Workshops (1)*, volume 7583 of *LNCS*, pages 433–442. Springer, 2012.

[3] N. Bird, O. Masoud, N. Papanikolopoulos, and A. Isaacs. Detection of Loitering Individuals in Public Transportation Areas. *IEEE Trans. on Intelligent Transportation Systems*, 6(2):167–177, June 2005.

[4] G. Doretto, T. Sebastian, P. H. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *J. Ambient Intelligence and Humanized Computing*, 2(2):127–151, 2011.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of CVPR*, pages 2360–2367, June 2010.

[6] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *Proc. of 10th IEEE Int. Workshop on PETS*, 2007.

[7] M. Hirzer, P. Roth, M. KÃűstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV 2012*, volume 7577 of *LNCS*, pages 780–793. Springer Berlin Heidelberg, 2012.

[8] R. Layne, T. M. Hospedales, and S. Gong. Towards person identification and re-identification with attributes. In *ECCV Workshops (1)*, volume 7583 of *LNCS*, pages 402–412. Springer, 2012.

[9] G. Pons-Moll, L. Leal-Taixé, T. Truong, and B. Rosenhahn. Efficient and robust shape matching for model based human motion capture. In *Proc. of ICPR*, DAGM'11, pages 416–425, Berlin, Heidelberg, 2011. Springer-Verlag.

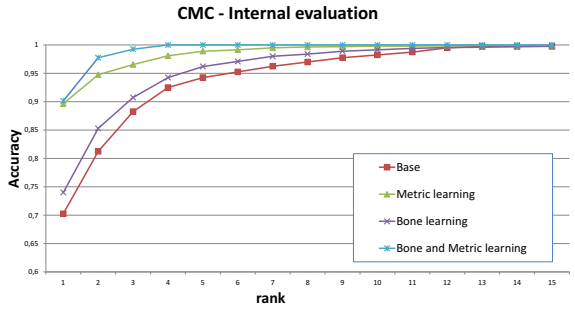
[10] M. Salzmann and R. Urtasun. Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In *Proc. of CVPR*, pages 647 –654, june 2010.

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. of CVPR*, pages 1297 –1304, june 2011.

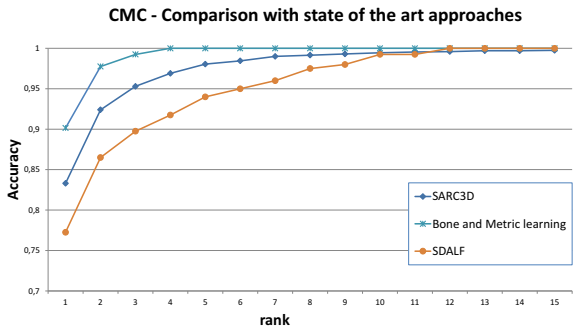
[12] J. Taylor, J. Shotton, T. Sharp, and A. W. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proc. of CVPR*, pages 103–110, 2012.

[13] M. Weber, M. BÃd'uml, and R. Stiefelhagen. Part-based Clothing Segmentation for Person Retrieval. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, page 6, Aug. 2011.

[14] X.-S. Yang and S. Deb. Engineering optimisation by cuckoo search. *Int. J. Mathematical Modelling and Numerical Optimisation*, 1:330–343, 2010.



(a)



(b)

Figure 5: CMC of the method on the Kinect dataset, showing a) the contribution of the metric and bone fragmentation learning and b) the improvement on two state of the art techniques