Re-identification with RGB-D Sensors

Igor Barros Barbosa[1,3], Marco Cristani[1,2], Alessio Del Bue[1],
Loris Bazzani[1], and Vittorio Murino[1]

[1] Pattern Analysis and Computer Vision (PAVIS),
Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163 Genova, Italy
[2] Dipartimento di Informatica, University of Verona,
Strada Le Grazie 15, 37134 Verona, Italy
[3] Université de Bourgogne, 720 Avenue de lEurope, 71200 Le Creusot, France

Abstract. People re-identification is a fundamental operation for any multi-camera surveillance scenario. Until now, it has been performed by exploiting primarily appearance cues, hypothesizing that the individuals cannot change their clothes. In this paper, we relax this constraint by presenting a set of 3D soft-biometric cues, being insensitive to appearance variations, that are gathered using RGB-D technology. The joint use of these characteristics provides encouraging performances on a benchmark of 79 people, that have been captured in different days and with different clothing. This promotes a novel research direction for the re-identification community, supported also by the fact that a new brand of affordable RGB-D cameras have recently invaded the worldwide market.

Keywords: Re-identification, RGB-D sensors, Kinect.

1 Introduction

The task of person re-identification (re-id) consists in recognizing an individual in different locations over a set of non-overlapping camera views. It represents a fundamental task for heterogeneous video surveillance applications, especially for modeling long-term activities inside large and structured environments, such as airports, museums, shopping malls, etc. In most of the cases, re-id approaches rely on appearance-based only techniques, in which it is assumed that individuals do not change their clothing within the observation period [1–3]. This hypothesis represents a very strong restriction, since it constraints re-id methods to be applied under a limited temporal range (reasonably, in the order of minutes).

In this paper we remove this restriction, presenting a new approach of person re-id that uses soft biometrics cues as features. In general, soft biometrics cues have been exploited in different contexts, either to aid facial recognition [4], used as features in security surveillance solutions [5, 6] or also for person recognition under a bag of words policy [7]. In [4] soft biometrics cues are the size of limbs, which were manually measured. The approaches in [5–7] are based on data coming from 2D cameras and extract soft biometrics cues such as gender, ethnicity, clothing, etc.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 433–442, 2012.

At the best of our knowledge, 3D soft biometric features for re-identification have been employed only in [4], but in that case the scenario is strongly supervised and needs a complete cooperation of the user to take manual measures. In contrast, a viable soft biometrics system should mostly deal with subjects without requiring strong collaboration from them, in order to extend its applicability to more practical scenarios.

In our case, the cues are extracted from range data which are computed using RGB-D cameras. Recently, novel RGB-D camera sensors as the *Microsoft Kinect* and *Asus Xtion PRO*, both manufactured using the techniques developed by PrimeSense [8], provided to the community a new method of acquiring depth information in a fast and affordable way. This drove researchers to use RGB-D cameras in different fields of applications, such as pose estimation [9] and object recognition [10], to quote a few. In our opinion, re-id can be extended to novel scenarios by exploiting this novel technology, allowing to overcome the constraint of analyzing people that do not change their clothes.

In particular, our aim is to extract a set of features computed directly on the range measurements given by the sensor. Such features are related to specific anthropometric measurements computed automatically from the person body. In more detail, we introduce two distinct subsets of features. The first subset represents cues computed from the fitted skeleton to depth data i.e. the Euclidean distance between selected body parts such as legs, arms and the overall height. The second subset contains features computed on the surface given by the range data. They come in the form of geodesic distances computed from a predefined set of joints (e.g. from torso to right hip). This latest measure gives an indication of the curvature (and, by approximation, of the size) of specific regions of the body.

After analyzing the effectiveness of each feature separately and performing a pruning stage aimed at removing not influent cues, we studied how such features have to be weighted in order to maximize the re-identification performance. We obtained encouraging re-id results on a pool of 79 people, acquired under different times and across intervals of days. This promotes our approach and in general the idea of performing re-id with 3D soft biometric cues extracted from RGB-D cameras.

The remaining of the paper is organized as follows. Section 2 briefly presents the re-identification literature. Section 3 details our approach followed by Section 4 that shows experimental results. Finally, Section 5 concludes the paper, envisaging some future perspectives.

2 State of the Art

Most of the re-identification approaches build on appearance-based features [1, 11, 3] and this prevents from focusing on re-id scenarios where the clothing may change. Few approaches constrain the re-id operative conditions by simplifying the problem to temporal reasoning. They actually use the information on the layout distribution of cameras and the temporal information in order to prune away some candidates in the gallery set [12].

The adoption of 3D body information in the re-identification problem was first introduced by [13] where a coarse and rigid 3D body model was fitted to different pedestrians. Given such 3D localization, the person silhouette can be related given the different orientations of the body as viewed from different cameras. Then, the registered data are used to perform appearance-based re-identification. Differently, in our case we manage genuine soft biometric cues of a body which is truly non-rigid and also disregarding an appearance based approach. Such possibility is given by nowadays technology that allows to extract reliable anatomic cues from depth information provided by a range sensor.

In general, the methodological approach to re-identification can be divided into two groups: learning-based and direct strategies. Learning based methods split a re-id dataset into two sets: training and test [1, 3]. The training set is used for learning features and strategies for combining features while the test dataset is used for validation. Direct strategies [11] are simple feature extractors. Usually, learning-based strategies are strongly time-consuming (considering the training and testing steps), but more effective than direct ones. Under this taxonomy, our proposal can be defined as a learning-based strategy.

3 Our Approach

Our re-identification approach has two distinct phases. First, a particular signature is computed from the range data of each subject. Such signature is a composition of several soft biometric cues extracted from the depth data acquired with a RGB-D sensor. In the second phase, these signatures are matched against the test subjects from the gallery set. A learning stage, computed beforehand, explains how each single feature has to be weighted when combined with the others. A feature with high weight means that it is useful for obtaining good re-identification performances.

3.1 First Stage: Signature Extraction

The first step processes the data acquired from a RGB-D camera such as the Kinect. In particular, this sensor uses a structured light based infrared patterns [8] that illuminates the scene/objects. Thus the system obtains a depth map of the scene by measuring the pattern distortion created by the 3D relief of the object. When RGB-D cameras are used with the *OpenNI* framework [14], it is possible to use the acquired depth map to segment & track human bodies, estimate the human pose, and perform metric 3D scene reconstruction. In our case, the information used is given by the segmented point-cloud of a person, the positions of the fifteen body joints and the estimation of the floor plane. Although the person depth map and pose are given by the *OpenNI*software libraries, the segmentation of the floor required an initial pre-processing using RANSAC to fit a plane to the ground. Additionally, a mesh was generated from the person point cloud using the“Greedy Projection” method [15].

Before focusing on the signature extraction, a preliminary study has been performed by examining a set of 121 features on a dataset of 79 individuals, each captured in 4 different days (see more information on the dataset in Sec. 4). These features can be partitioned in two groups: the first contains the *skeleton-based features*, i.e., those cues which are based on the exhaustive combination of distances among joints, distances between the floor plane and all the possible joints. The second group contains the *Surface-based features*, i.e., the geodesic distances on the mesh surface computed from different joints pairs. In order to determine the most relevant features, a feature selection stage evaluates the performance on the re-identification task of each single cue, one at a time, independently. In particular, as a measure of the re-id accuracy, we evaluated the normalized area under curve (nAUC) of the cumulative matching curve (CMC) discarding those features which resulted equivalent to perform a random choice of the correct match (see more information on these classification measures on Sec. 4).

The results after such pruning stage was a set of 10 features:

- **Skeleton-based features**
  - d_1: Euclidean distance between floor and head
  - d_2: Ratio between torso and legs
  - d_3: Height estimate
  - d_4: Euclidean distance between floor and neck
  - d_5: Euclidean distance between neck and left shoulder
  - d_6: Euclidean distance between neck and right shoulder
  - d_7: Euclidean distance between torso center and right shoulder
- **Surface-based features**
  - d_8: Geodesic distance between torso center and left shoulder
  - d_9: Geodesic distance between torso center and left hip
  - d_{10}: Geodesic distance between torso center and right hip

Some of the features based on the distance from the floor are illustrated in Fig. 1 together with the joints localization on the body. In particular, the second feature (ratio between torso and legs) is computed according to the following equation:

$$d_2 = \frac{mean(d_5 + d_6)}{mean(d_{floorLhip} + d_{floorRhip})} \cdot (d_1)^{-1} \tag{1}$$

The computation of the (approximated) geodesic distances, i.e., *Torso to left shoulder, torso to left hip* and *torso to right hip*, is given by the following steps. First, the selected joints pairs, which are normally not lying onto the point cloud, are projected towards the respective closest points in depth. This generates a starting and ending point on the surface where it is possible to initialize an $A^\star$ algorithm computing the minimum path over the point cloud (Fig. 2). Since the torso is usually recovered by the RGB-D sensor with higher precision, the computed geodesic features should be also reliable.

As a further check on the 10 selected features, we verified the accuracy by manually measuring the features on a restricted set of subjects. At the end, we found out that higher precision was captured especially in the features related to

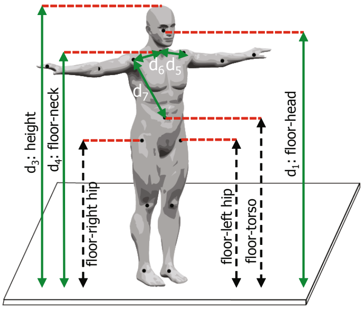

Fig. 1. Distances employed for building the soft-biometric features (in black), and some of the soft biometric features (in green). It is important to notice that the joints are not localized in the outskirt of the point-cloud, but, in most of the cases, in the proximities of the real articulations of the human body.

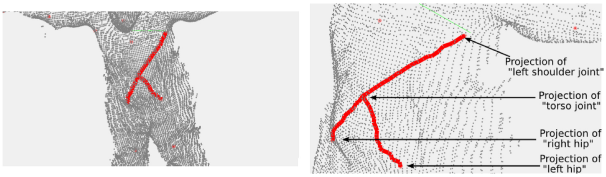
Projection of "left shoulder joint"
Projection of "torso joint"
Projection of "right hip"
Projection of "left hip"

Fig. 2. Geodesic features: the red line represents the path found by $A^{\star}$ between *torso to left shoulder, torso to left hip* and *torso to right hip*

the height $(d_1, ..., d_4)$, while other features were slightly more noisy. In general, all these features are well-suited for an indoor usage, in which people do not wear heavy clothes that might hide the human body aspects.

3.2 Second Stage: Signature Matching

This section illustrates how the selected features can be jointly employed in the re-id problem. In the literature, a re-id technique is usually evaluated considering two sets of personal ID signatures: a gallery set A and a probe set B.

The evaluation consists in associating each ID signature of the probe set B to a corresponding ID signature in the gallery set A. For the sake of clarity, let us suppose to have N different ID signatures (each one representing a different individual, so N different individuals) in the probe set and the same occurs in the gallery set. All the N subjects in the probe are present in the gallery. For evaluating the performance of a re-id technique, the most used measure is the Cumulative Matching Curve (CMC) [1], which models the mean probability that whatever probe signature is correctly matched in the first T ranked gallery individuals, where the ranking is given by evaluating the distances between ID signatures in ascending order.

In our case, each ID signature is composed by F features (in our case, $F = 10$), and each feature has a numerical value. Let us then define the distance between corresponding features as the squared difference between them. For each feature, we obtain a $N \times N$ distance matrix. However such matrix is biased towards features with higher measured values leading to a problem of heterogeneity of the measures. Thus, if a feature such as the height is measured, it would count more w.r.t. other features whose range of values is more compact (e.g. the distance between neck and left shoulder). To avoid this problem, we normalize all the features to a zero mean and unitary variance. We use the data from the gallery set to compute the mean value of each feature as well as the feature variance.

Given the normalized $N \times N$ distance matrix, we now have to surrogate those distances into a single distance matrix, obtaining thus a final CMC curve. The naive way to integrate them out would be to just average the matrices. Instead, we propose to utilize a weighted sum of the distance matrices. Let us define the set of weight w_i for $i = 1, ..., F$ that represents the importance of the $i-$th feature: the higher the weight, the more important is the feature. Since tuning those weights is usually hard, we propose a *quasi-exhaustive* learning strategy, i.e., we explore the weight space (from 0 to 1 with step 0.01) in order to select the weights that maximize the nAUC score. In the experiments, we report the values of those weights and compare this strategy with the average baseline.

4 Experiments

In this section, we describe first how we built the experimental dataset and how we formalised the re-id protocol. Then, an extensive validation is carried forward over the test dataset in different conditions.

4.1 Database Creation

Our dataset is composed by four different groups of data. The first "Collaborative" group has been obtained by recording 79 people with a frontal view, walking slowly, avoiding occlusions and with stretched arms. This happened in an indoor scenario, where the people were at least 2 meters away from the camera. This scenario represents a collaborative setting, the only one that we considered in these experiments. The second ("Walking") and third ("Walking2") groups of data are composed by frontal recordings of the same 79 people walking normally while entering the lab where they normally work. The fourth group ("Backwards") is a back view recording of the people walking away from the lab. Since all the acquisitions have been performed in different days, there is no guarantee that visual aspects like clothing or accessories will be kept constant. Figure 3 shows the computed meshes from different people during the recording of the four different sessions, together with some statistics about the collected features.

From each acquisition, a single frame was automatically selected for the computation of the biometric features. This selection uses the frame with the best

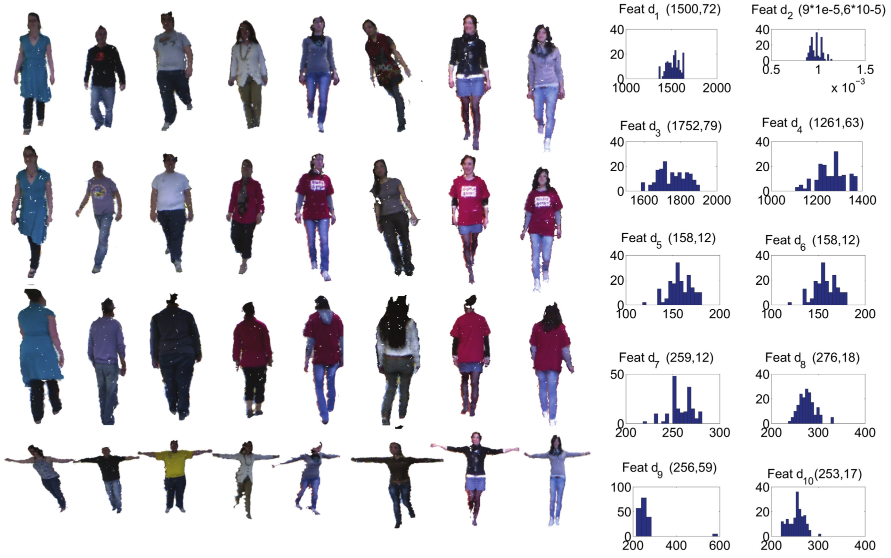

Fig. 3. Illustration of the different groups in the recorded data, rows from top to bottom: "Walking", "Walking2", "Backwards" and "Collaborative". Note that people changed their clothings during the acquisitions in different days. On the right, statistics of the "Walking" dataset: for each feature, the histogram is shown; in the parenthesis, its mean value (in cm, except d_2) and standard deviation.

confidence of tracked skeleton joints[1], which is closest to the camera and it was not cropped by the sensors fields of view. This represents the frame with the highest joints tracking confidence which in most of the cases was approximately 2.5 meters away from the camera.

After that, the mesh for each subject was computed and the 10 soft biometric cues have been extracted using both skeleton and geodesics information.

4.2 Semi-cooperative re-id

Given the four datasets, we have built a semi-collaborative scenario, where the gallery set was composed by the ID signatures of the "Collaborative" setting, and the test data was the "Walking 2" set. The CMCs related to each feature are portrayed in Fig. 4: they show how each feature is able to capture discriminative information of the analyzed subjects. Fig. 5 shows the normalized AUC of each features. Notice that the features associated to the height of the person are very meaningful, as so the ratio between torso and legs.

The results of Fig. 5 highlights that the nAUC over the different features spans from 52.8% to 88.1%. Thus, all of them contributes to have better re-identification

[1] Such confidence score is a byproduct of the skeleton fitting algorithm.

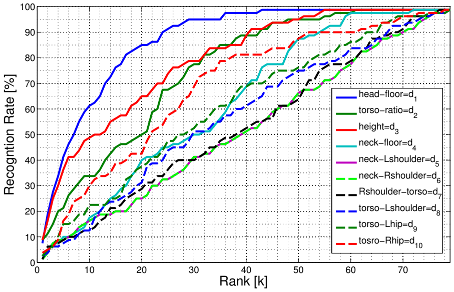

Fig. 4. Single-feature CMCs — "Collaborative" VS "Walking 2" (best viewed in colors)

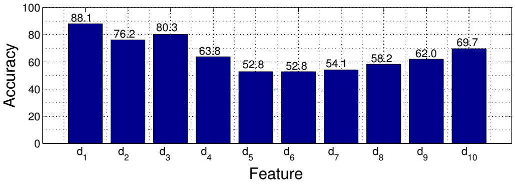

Fig. 5. Area under the curve for each feature (the numbering here follows the features enumeration presented in Sec. 3) —"Collaborative" VS "Walking 2". The numbers over the bars indicate the numerical nAUC values of the different features.

results. To investigate how their combination helps in re-id, we exploit the learning strategy proposed in Sec. 3.2. Such weights w_i are learned once using a different dataset than the one used during testing. The obtained weights are: $w_1 = 0.24, w_2 = 0.17, w_3 = 0.18, w_4 = 0.09, w_5 = 0.02, w_6 = 0.02, w_7 = 0.03, w_8 = 0.05, w_9 = 0.08, w_{10} = 0.12$. The weights mirrors the nUAC obtained for each feature independently (Fig. 5): the most relevant ones are d_1 (Euclidean distance between floor and head), d_2 (Ratio between torso and legs), d_3 (Height estimate), and d_{10} (Geodesic distance between torso center and right hip). In Fig. 6, we compare this strategy with a baseline: the average case where $w_i = 1/F$ for each i. It is clear that the learning strategy gives better results (nAUC= 88.88%) with respect to the baseline (nAUC= 76.19%) and also the best feature (nAUC= 88.10%) that correspods to d_1 in Fig. 5. For the rest of the experiments the learning strategy is adopted.

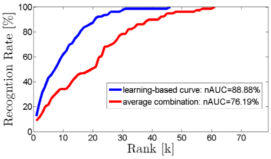

Fig. 6. Compilation of final CMC curves —"Collaborative" - "Walking 2"

4.3 Non-cooperative re-id

Non-cooperative scenarios consist of the “walking”, “walking2” and “backwards” datasets. We generate different experiments by combining cooperative and non-cooperative scenarios as gallery and probe sets. Table 1 reports the nAUC score given the trials we carried out. The non-cooperative scenarios gave rise to higher performances than the cooperative ones. The reason is that, in the collaborative acquisition, people tended to move in a very unnatural and constrained way, thus originating biased measurements towards a specific posture. In the non-cooperative setting this did not clearly happen.

Table 1. nAUC scores for the different re-id scenarios

Gallery	**Probe**	**nAUC**
Collab.	Walking	90.11 %
Collab.	Walking 2	88.88 %
Collab.	Backwards	85.64 %
Walking	Walking 2	91.76 %
Walking	Backwards	88.72%
Walking 2	Backwards	87.73 %

5 Conclusions

In this paper, we presented a person re-identification approach which exploits soft-biometrics features, extracted from range data, investigating collaborative and non-collaborative settings. Each feature has a particular discriminative expressiveness with height and torso/legs ratio being the most informative cues. Re-identification by 3D soft biometric information seems to be a very fruitful research direction: other than the main advantage of a soft biometric policy, i.e., that of being to some extent invariant to clothing, many are the other reasons: from one side, the availability of precise yet affordable RGB-D sensors encourage the study of robust software solutions toward the creation of real surveillance system. On the other side, the classical appearance-based re-id literature is characterized by powerful learning approaches that can be easily embedded in the 3D situation. Our research will be focused on this last point, and on the creation of a larger 3D non-collaborative dataset.

References

1. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
2. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)

3. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 649–656. IEEE (2011)
4. Velardo, C., Dugelay, J.-L.: Improving identification by pruning: a case study on face recognition and body soft biometric. Eurecom, Tech. Rep. EURECOM+3593 (January 2012)
5. Wang, Y.-F., Chang, E.Y., Cheng, K.P.: A video analysis framework for soft biometry security surveillance. In: Proceedings of the third ACM International Workshop on Video Surveillance & Sensor Networks, VSSN 2005, pp. 71–78 (2005)
6. Demirkus, M., Garg, K.: Automated person categorization for video surveillance using soft biometrics. In: Proc of SPIE, Biometric Technology (2010)
7. Dantcheva, A., Dugelay, J.-L., Elia, P.: Person recognition using a bag of facial soft biometrics (BoFSB). In: 2010 IEEE International Workshop on Multimedia Signal Processing, pp. 511–516. IEEE (October 2010)
8. Freedman, B., Shpunt, A., Machline, M., Ariel, Y.: US Patent - US2010/0118123 (2010)
9. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011, pp. 1297–1304. IEEE (June 2011)
10. Bo, L., Lai, K., Ren, X., Fox, D.: Object recognition with hierarchical kernel descriptors. In: CVPR 2011, pp. 1729–1736. IEEE (June 2011)
11. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: British Machine Vision Conference, BMVC (2011)
12. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Comput. Vis. Image Underst. 109(2), 146–162 (2008)
13. Baltieri, D., Vezzani, R., Cucchiara, R.: SARC3D: A New 3D Body Model for People Tracking and Re-identification. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part I. LNCS, vol. 6978, pp. 197–206. Springer, Heidelberg (2011)
14. OpenNI (February 2012) Openni framework@ONLINE, `http://www.openni.org/`
15. Marton, Z.C., Rusu, R.B., Beetz, M.: On Fast Surface Reconstruction Methods for Large and Noisy Datasets. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, May 12-17 (2009)