2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

A Feature-based Approach to People Re-Identification using Skeleton Keypoints

Matteo Munaro, Stefano Ghidoni, Deniz Tartaro Dizmen and Emanuele Menegatti

Abstract— **In this paper we propose a novel methodology for people re-identification based on skeletal information. Features are evaluated on the skeleton joints and a highly distinctive and compact feature-based signature is generated for each user by concatenating descriptors of all visible joints. We compared a number of state-of-the-art 2D and 3D feature descriptors to be used with our signature on two newly acquired public datasets for people re-identification with RGB-D sensors. Moreover, we tested our approach against the best re-identification methods in the literature and on a widely used public video surveillance dataset. Our approach proved to be robust to strong illumination changes and occlusions. It achieved very high performance also on low resolution images, overcoming state-of-the-art methods in terms of recognition accuracy and efficiency. These features make our approach particularly suited for mobile robotics.**

I. Introduction

People re-identification is the capability of associating a new observation of a person to others made in the past. Distinguishing the different persons that are in the environment is a high-level capability that is crucial in several fields including service robotics, intelligent video surveillance systems and smart environments. Strong efforts have been spent by the research community to improve performance and reliability of re-identification algorithms, and several different techniques have been developed.

People re-identification in images is addressed observing three main characteristics: color, texture and shape, either considered singularly or mixed together. Color undergoes clear changes from person to person, and is usually measured by means of global or partial histograms. A color-based state-of-the-art approach divides the body of each target into smaller parts and evaluates multiple histograms, one for each part [1], [2]. This method is simple and effective, but suffers from two main flaws: it fails upon strong illumination changes, and it is a global (or semi-global) method that is not able to describe the target in detail.

Texture-based and shape-based approaches usually make use of local features: they exploit descriptors evaluated on a set of keypoints to generate the signature of a target. Performance are therefore strongly related to the characteristics of the set of descriptors selected, including the capability of the keypoint detector to select stable features. This approach is widely used in the literature [3], [4], [5] thanks to its superior capability of providing a detailed description of each target; moreover, it overcomes the two main drawbacks of the color-based approach previously discussed. Such approach has also been used together with histograms [6].

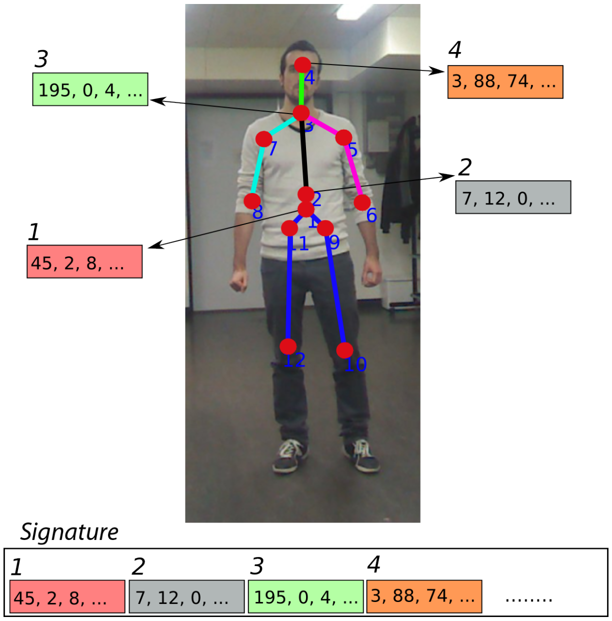


Fig. 1. An overview of our approach to re-identification. Feature descriptors are computed around human skeleton joints, which serve as keypoints. These descriptors are then concatenated to obtain the whole person signature.

Very recently, computer vision for robotics was revolutionized by the introduction of affordable high-resolution three-dimensional sensors, that generate color point clouds instead of images. This had a strong impact on a number of applications, including people re-identification: for example, approaches based on three-dimensional features were developed. This new type of features might include information about both color and shape, which can provide superior performance over standard 2D features; however, noise affecting the location of the point cloud elements is usually not negligible for low-cost sensors like the Microsoft Kinect: in this case, shape descriptors can provide performance worse than expected, thus global approaches are usually preferred [7], [8], [9].

In this paper we present a novel three-dimensional feature-based re-identification system capable of providing superior performance while preserving computational efficiency, that is crucial for real-time applications. Our approach is particularly suited for robotics applications, that have strong requirements on computational load and processing time. Moreover, it makes use of a sensor that is often available

The authors are with the Intelligent Autonomous Systems Laboratory (IAS-Lab) at the University of Padua, Padua, IT. {munaro, ghidoni, emg}@dei.unipd.it.

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

on robotic platforms, and is already used by a large number of people in the research community. We demonstrate that our system is able to exploit the high amount of information provided by RGB-D sensors to achieve better performance with respect to other state-of-the-art approaches.

As discussed above, several feature-based approaches presented in the literature compute the signature of a target by finding keypoints and evaluating the features, either on 2D or 3D data, which is a general-purpose technique that works for any type of target. Differently, we base our approach on the assumption that the target is a human, whose body has a well-known structure. Building on such hypothesis, we exploit a skeletal tracker to evaluate the body pose, and use this information to guide the keypoint selection, as shown in Fig. 1, and compute the proposed *Skeleton-based Person Signature* (SPS).

A small number of points are found at specific body positions: head, neck, shoulders, elbows, hands, ankles, knees, feet. Only 15 or 20[1] keypoints need to be evaluated and compared for every target: this dramatically reduces the computational load and the amount of memory for storing target models. The small number of keypoints is compensated by their high stability: since their positions come from a high-level knowledge of the target shape, their variance is extremely low compared to the typical values achievable with low-level keypoint detectors.

We thoroughly studied and tested our novel skeleton-based approach. Several state-of-the-art 2D and 3D features were considered, leading to a mixed 3D-2D approach, which provides the best performance: the skeleton is evaluated from 3D data, while the signature is obtained from 2D descriptors. The presented system is meant to provide the best performance on high-resolution RGB-D data; however, benchmark datasets available for testing re-identification algorithms are often made of low-resolution 2D images. We collected the IAS-Lab RGBD-ID dataset, a novel dataset for testing our approach, acquired using a Microsoft Kinect installed on a robot, so that people are seen from the typical perspective used in robotics applications. This dataset highlights the system performance in the targeted working conditions. We also performed further experiments to validate our algorithms with respect to previous approaches in the literature: to do so, we chose the CAVIAR4REID video-surveillance dataset, that is widely used. Since our approach is based on a skeletal tracker, we added the joints to the dataset, applied our method, and compared our approach to others in the literature, showing it advances the state-of-the-art also in this case. As a further contribution to the research community, we released both the new dataset and the skeletal joints of the CAVIAR4REID dataset.

Our re-identification approach was designed to be accurate and fast at the same time: for this reason, it is suitable for being included in real-time people tracking systems used in robotics, offering a superior performance with respect to appearance classifiers commonly exploited in tracking systems.

The paper is organized as follows: in Sec. II the re-identification system based on our novel approach is presented in its main components: skeletal tracker (Sec. II-A), exploited features (Sec. II-B) and methods for matching targets (Sec. II-C). In Sec. III, experiments run for assessing the performance and comparing our approach to the state-of-the-art are discussed. In Sec. IV, some final remarks about the presented systems are reported.

II. System overview

The proposed approach to re-identification characterizes each target based on several feature vectors, evaluated on a set of keypoints. We call *keypoints* the salient points taken as reference to characterize the target, i.e. the locations where the target is observed. A *descriptor* is a vector that describes the characteristics (like color, shapes, etc.) of a keypoint neighborhood, and is the output of a feature extraction algorithm. The number and type of characteristics taken into account, as well as the size of the neighborhood, depend on the specific feature. Finally, the *signature* is the data that describes the whole target, normally obtained by concatenating all descriptors in an organized way.

The two main components of our approach are the skeletal tracker, that determines the keypoint location, and the type of feature chosen for characterizing them. Both elements have a strong impact on the re-identification performance, and have been deeply studied in this work.

A. Skeletal Tracker

The stability of the keypoints is extremely important to obtain similar descriptors in different observations of the same target, thus ensuring a reliable re-identification. Dealing with human people is very different than working with objects, since the human body is very flexible and deformable, and usually presents a number of different texture and shape patterns on the different body parts.

To properly describe the body shape and pose we adopted a skeletal tracker, that is an algorithm capable of understanding how the human body is placed; its output is a set of points, called *joints*, that represent a small subset of the real joints of a body.

In our work we exploited two skeletal trackers, that are the most commonly used in computer vision and robotics: one is provided by Microsoft as a complement to its Kinect sensor and will be called KST (Kinect Skeletal Tracker) [10]; the other is included in the OpenNI SDK[2], is implemented inside the Nite Middleware[3] and will be called NST (Nite Skeletal Tracker). Both trackers work on 3D data, which ensures an accurate description of the body shape and superior performance over the 2D-based approaches.

The KST tracks 20 joints of the human body at 30 fps. It allows to obtain a precise estimation from a single depth image, but it assumes that people are facing the camera, thus

[1] Depending on how many joints are estimated by the skeletal tracking algorithm.

[2] `http://www.openni.org`.

[3] `http://www.openni.org/files/nite/`.

the skeleton is flipped left-right if the person is seen from the back side. Instead, the NST tracks 15 joints (no hands and feet) and exploits the information from multiple frames to improve the tracking performance. Unlike KST, it allows to track skeletons also for people seen from the back side and it can be used with the Robot Operating System (ROS) [11].

Depending on the quality of the segmented target and the level of occlusion, the skeletal trackers might not detect all the joints: in this case, some of them are marked as non tracked, meaning their location is not reliable, but they are anyway part of the whole skeleton. This situation is rather common for both skeletal trackers, even though they show a different behavior in such case: from our experience, while KST provides reasonable skeletons also when some joints are not tracked, NST often poorly estimates the whole skeleton when some joints are not visible. KST provides bad skeleton estimations mainly when the person is partially occluded or out of the image to the right or left side, or it is farther than 3.5 m from the sensor. In Fig. 2 we report two examples of these situations together with a case of good estimation.

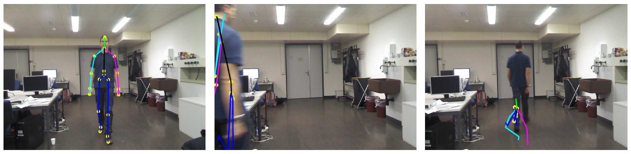

(a) Frontal (b) Out of the image (c) Too far

Fig. 2. Example of good skeleton estimation (a) and two situations in which KST estimates wrong skeletons, caused by the target being only partially visible (b) or too far from the sensor (c).

B. Descriptor Evaluation

Once keypoints have been located by the skeletal tracker, features are exploited to characterize their appearance in terms of texture, color and three-dimensional shape. Features considered in this work were taken from the most recent literature in the fields of robotics and computer vision; both 2D and 3D variants were considered, because they offer advantages and disadvantages. Our system therefore relies on 3D information for the evaluation of the body pose, and either on 2D or 3D data for evaluating descriptors and generate the signature. The sensor exploited for the experiments provides both 2D images and 3D point clouds of the scene. Calibration data is also available, which can be used for mapping correspondences between given locations of the point cloud on the 2D image. This knowledge is exploited to enable the use of 2D features on 3D keypoints.

Generally speaking, 3D features are often more computationally expensive to evaluate and noisy, because the additional noise source affecting the point position in a point cloud is not negligible when low-cost sensors like the Microsoft Kinect are exploited for acquisition. On the other hand, 3D features are obviously highly descriptive and provide richer information. For a more detailed discussion we refer to the literature presenting keypoint and feature detectors that is reported later in this section.

Dealing with features, the radius on which they are evaluated is an important parameter that has a strong impact on the performance. For 3D features, such radius R is a metric value which can be chosen once and kept constant, while in the image domain the feature radius r is measured in pixels and cannot be fixed because the size of a person in the image varies depending on its 3D position. A great advantage offered by our 3D-based approach is the capability of relating r expressed in pixels to R expressed in world coordinates and fixed for all targets in all views: this sensibly enhances the feature stability, because the same volume is always considered for evaluating the features. This is possible because the relationship between the two radii r and R can be expressed by:

$$r = f \times \frac{R}{z}, \tag{1}$$

where z is the distance of the target to the sensor plane and f is the sensor focal length.

The implementations chosen for evaluating the features discussed in the following are the standard ones provided by the widely diffused OpenCV [12] and PCL [13] libraries, using default parameters when available. 2D descriptors used for our tests are: SIFT [14], SURF [15], BRIEF [16], ORB [17] and FREAK [18]. Among the 3D features, the following ones were chosen for testing our approach: PFH [19] and its variation including color information, called PFHRGB; FPFH [20], SHOT [21], SHOTRGB [22].

The features we employed for re-identification are widely used in the robotics field, but are targeted to rigid objects. Instead the human body is deformable: this is considered in our approach by means of the skeletal tracker, but the feature vector evaluation process considers each joint as a part of a rigid object. This approximation is reasonable for targets acquired at low or medium resolution, and leads to good results. Warping the target point cloud to a standard pose [23] could further improve the re-identification results.

C. Matching Methods

The re-identification task is achieved by comparing the signatures of each target found in the test frame with those found in the training frames; the best-matching one is finally selected. We propose two ways of considering the skeleton joints in the matching phase: one considers tracked joints only, the other all of them.

1) Tracked Joints Matching Method: The first proposed method works by matching the reliable joints only, i.e. those that are in the tracked state in the test frame, and is therefore named TJ (Tracked Joints). This method achieves high performance (as it will be discussed in Sec. III) because it relies on visible and stable joints only. Additionally, this method deals with almost all frames in each dataset, because it does not require the complete skeleton to process a frame, but rather, is able to select the part that has been reliably detected. The only exception occurs when the target is partially outside the field of view, because the whole skeleton is poorly detected.

2) All Joints Matching Method: The second approach is called AJ (All Joints) because it considers also the unstable keypoints – recall that all of them are always located by both skeletal trackers considered in this work. The performance of this approach is lower with respect to TJ because unreliable keypoints are considered in the matching. This approach is nevertheless proposed to test our system on the CAVIAR4REID dataset, in which training and testing frames are often acquired from different viewpoints. This represents a special case, because some joints that are detected in the test frames are never seen in the training set, and would therefore never be matched, thus reducing the number of joints actually usable: the AJ matching method is therefore used for recovering this situation.

When the AJ matching is used, an additional operation is performed. If a joint is not tracked, its descriptor is replaced with the descriptor computed at its symmetric joint, if this one is tracked. This change relies on the assumption that descriptors at symmetric joints are similar, thus this replacement can increase the recognition performance. For instance, if the right hand is not tracked while the left hand is tracked, the descriptor around the left hand is used in place of that of the right hand. This obviously applies to all symmetric joint couples.

D. Skeleton-based Person Signature

The signature we propose for describing a target is called Skeleton-based Person Signature (SPS). It is built according to the matching method adopted, therefore two versions of the signature were developed: one evaluated only on the tracked keypoints, the other on all of them. In the first case we first define the feature tracking indicator for the i-th joint J_i on the k-th target T_k as:

$$I\left(J_i, T_k\right) = \begin{cases} 1 & \text{if } i\text{-th joint of frame } k \text{ is tracked} \\ 0 & \text{otherwise} \end{cases} , \quad (2)$$

where $i \in [0, \ldots, N-1]$ and N is the number of joints considered; note that KST and NST automatically provide $I\left(J_i, T_k\right)$. The signature on tracked joints $\mathbf{SPS}^{\mathrm{TJ}}$ of target k is obtained as:

$$\mathbf{SPS}_k^{\mathrm{TJ}} = \bigcup_{i=0}^{N-1} \left\{ D\left(J_i, T_k\right) : I\left(J_i, T_k\right) = 1 \right\} , \quad (3)$$

where $D\left(J_i, T_k\right)$ is the descriptor obtained evaluating the selected feature on the i-th joint for target k. Following (3), the signature is therefore obtained concatenating the descriptors of all tracked joints. The second version of the signature, built considering all joints, is analogously defined by simply omitting the feature tracking indicator:

$$\mathbf{SPS}_k^{\mathrm{AJ}} = \bigcup_{i=0}^{N-1} \left\{ D\left(J_i, T_k\right) \right\} . \quad (4)$$

E. Similarity metrics

The algorithms used at low level for comparing features are the standard methods commonly used, and already available in the OpenCV and PCL libraries we exploited for computing the features. In particular, two similarity metrics are exploited at this stage, depending on the data type of the feature descriptor, that can be composed either by real values, or binary ones. In the first case, the Euclidean distance is used, defined as:

$$d_{\mathrm{E}}\left(T_i, T_j\right) = \sqrt{\sum_{m=0}^{L-1} \left(\mathbf{SPS}_i(m) - \mathbf{SPS}_j(m)\right)^2} , \quad (5)$$

where $\mathbf{SPS}_i(m)$ is the m-th element of the signature, and $\mathbf{SPS}_i$ stands either for $\mathbf{SPS}_i^{\mathrm{TJ}}$ or $\mathbf{SPS}_i^{\mathrm{AJ}}$, depending on the type of signature to be considered; finally, L is the number of descriptors available in the signature, that can be smaller than the number of joints if $\mathbf{SPS}_i^{\mathrm{TJ}}$ is employed. The distance evaluated in (5) depends on the specific feature employed, but since only one type of feature is employed at a time, values are coherent, and can be used for the re-identification task.

In the second case, binary features provide descriptors that are bitstrings, and we compare them using the Hamming distance:

$$d_{\mathrm{H}}\left(T_i, T_j\right) = \sum_{m=0}^{L-1} \left[\mathbf{SPS}_i(m) - \mathbf{SPS}_j(m)\right]^2 . \quad (6)$$

F. Single-frame vs Multi-frame Re-Identification

As discussed in the previous section, re-identification means associating a target found in the current frame (test set) with others observed in the past, or a set of pre-recorded examples (training set). We already discussed how to classify a new test frame by matching it with the training frames.

When multiple frames of a person are available in the testing set, they can be jointly exploited to provide a more robust classification. In [2], authors propose a multi-shot modality which, for comparing M probe signatures of a given subject against N gallery signatures of another one, simply calculates all the possible $M \times N$ single-shot distances, and selects the smallest one. This approach does not offer an efficient scalability over the number of frames. The computational time grows over time since an increasing number of frames are considered for finding the best match.

Since our purpose is to use our re-identification approach in a real time scenario, we perform single-frame classification and adopt a voting scheme that associates each test sequence to the subject voted by the highest number of frames. This approach merges single-frame results into a sequence-wise result without adding further computational costs and leading to considerable improvements (10-30%) in the recognition rate, as we will see in Sec. III.

III. Experiments

For validating the approach we described in Sec. II, we performed a number of experiments on three different datasets, to assess system performance on different working conditions. We selected publically available datasets presenting different challenges related to the re-identification task. In particular, the *BIWI RGBD-ID dataset*[4] is targeted to people re-identification from a robot point of view when the training set is composed by many people.

The novel *IAS-Lab RGBD-ID dataset*[5] presents strong illumination changes because training and testing sets were acquired in different rooms; finally, the *CAVIAR4REID dataset*[6] is made of very low resolution images and contains occlusions, considerable pose changes between training and testing set and a high number of people since it is targeted to people re-identification in a video surveillance scenario.

For evaluation purposes, we compute *Cumulative Matching Characteristic (CMC) Curves* [24], which are commonly used for evaluating re-identification algorithms. For every k from 1 to the number of training subjects, these curves express the average person recognition rate computed when considering a classification to be correct if the ground truth person appears among the subjects who obtained the k best classification scores. The typical evaluation parameters for these curves are the *rank-1* recognition rate and the *normalized Area Under Curve* (nAUC), which is the integral of the CMC. In this work, the recognition rates are separately computed for every subject and then averaged to obtain the final recognition rate.

A. Performance analysis on the BIWI RGBD-ID dataset

The BIWI RGBD-ID dataset was originally targeted to long-term re-identification, thus people wear different clothes in the training video with respect to their two testing sequences. For our purposes, we used only the sequences in which people are wearing the same clothes. In particular, for every person, we exploited the testing video where the person is still as our training set and the testing video where the person is walking as our testing set. Thus, our training set was composed of 28 people and all the people in the testing set were also present in the training set.

In Fig. 3(a), we report the CMCs we obtained on this dataset with the $\mathbf{SPS}_k^{\mathrm{TJ}}$ signature computed exploiting the five 2D descriptors we tested and exploiting the TJ matching approach. The solid curves refer to a SPS which concatenates descriptors at all the 20 skeleton keypoints, while the dashed curves are obtained by removing from the signature the keypoints at wrists, hands, ankles and feet, that are the joints which are more often misplaced by the skeletal tracker. We labeled this approach *No Hands and Feet* (NHF). As it can be noticed, these two configurations lead to similar results, with SIFT obtaining the best rank-1 (65.67%) and nAUC (94.26%), followed in order by BRIEF, ORB and SURF. Unlike the others, the FREAK descriptor led to random results, with a nAUC around 50%, thus showing not to be discriminative for the re-identification task. These results are obtained performing re-identification on nearly all the frames where a skeleton is provided[7], with a number of tracked joints ranging from 10 to 20 for the solid curves. A complete analysis of the re-identification performance when varying the number of tracked joints (and thus of skeleton keypoints) is reported in Fig. 4. In particular, we report the rank-1 score when performing re-identification on only those frames for which *at least* J joints are tracked and when *exactly* J joints are tracked. As expected, this score increases when exploiting more skeleton keypoints in our SPS signature. However, the more joints are requested for computing the signature, the fewer frames are considered valid for performing re-identification. For understanding the trade-off between rank-1 and number of analyzed frames, in Fig. 4(b), we report the percentage of *valid* frames vs the number of tracked joints for our testing set from the BIWI RGBD-ID dataset. Moreover, Fig. 4(c) shows the number of frames in which each skeleton joint is tracked. It can be noticed that, if re-identification is only performed when all the 20 joints are tracked, only 37% of the frames with a skeleton are analyzed. That is the case reported in Fig. 3(b), where we can notice that rank-1 scores are considerably higher than when all frames are considered, and of Fig. 3(c), where 3D descriptors are computed at skeleton keypoints given by the NHF approach, because bad keypoint localization at hands and feet could easily lead to singularities in 3D descriptors. As expected, descriptors only encoding 3D local shape (PFH, FPFH, SHOT) achieve random performance, because of the strong noise affecting point clouds when consumer depth sensors are employed. Instead, descriptors encoding both shape and color achieve good performance, even if lower than those obtained with 2D descriptors. In particular, PFHRGB obtains the best rank-1 (48.49%) and nAUC (87.01%), while SHOTRGB stops at a rank-1 of 23.01% and a nAUC of 81.14%.

In Fig. 3(d), we compare the best 2D (SIFT) and 3D (PFHRGB) approaches of Fig. 4fig:BIWI(b) and (c) with other methods widely used in literature. In order to validate our choice of skeleton joints as keypoints, we also reported the CMC obtained when selecting keypoints with the standard SIFT keypoint detector and then matching the SIFT descriptors for re-identification, as often done for object recognition. It can be noticed how our approach to keypoint selection allows to obtain a rank-1 20% higher while also avoiding the feature matching step, which is the process of finding corresponding features among training and testing descriptors, thus saving a considerable amount of time, as we will see in Sec. III-E.

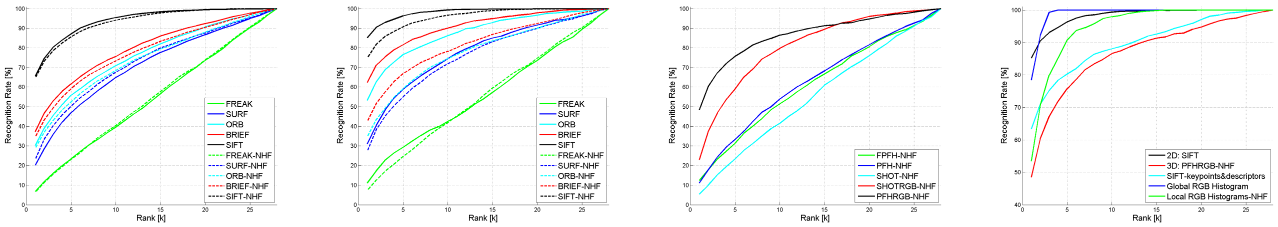
We also compare these approaches to the use of color histograms, which are highly used in people re-identification literature [1], [2], matched with the Bhattacharyya distance [25]. The best result is obtained by computing a global RGB histogram on all points belonging to a target, while a lower result is achieved when concatenating local RGB histograms extracted from each body part. However, such performance is still lower in terms of rank-1 score to our best SPS signature

[4] `http://robotics.dei.unipd.it/reid`.

[5] `http://robotics.dei.unipd.it/reid`.

[6] `http://www.lorisbazzani.info/code-datasets/caviar4reid`.

[7] We only discard those frames where the person is partially out of the image to the right or left or where the person is farther than 3.5 m because the skeleton is poorly estimated in these conditions.



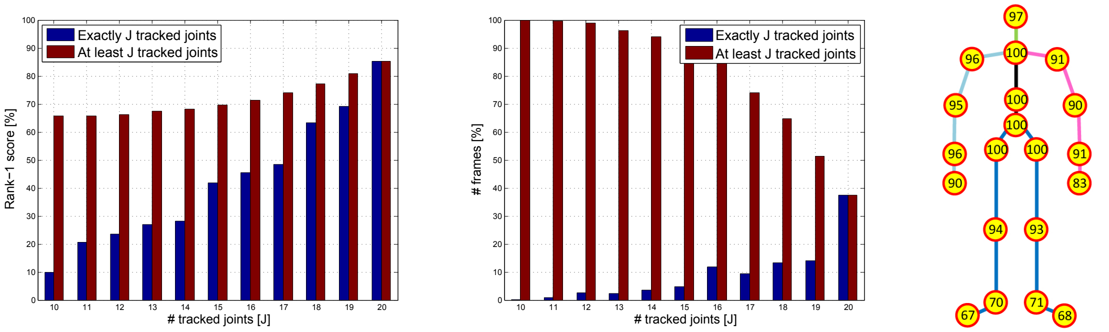
(a) 2D features - all frames

(b) 2D features - fully tracked skeletons

(c) 3D features - fully tracked skeletons

(d) Main tested approaches - fully tracked skeletons

Fig. 3. Re-identification results on the BIWI RGBD-ID dataset.



(a) Rank-1 score VS number of tracked joints

(b) Number of frames VS number of tracked joints

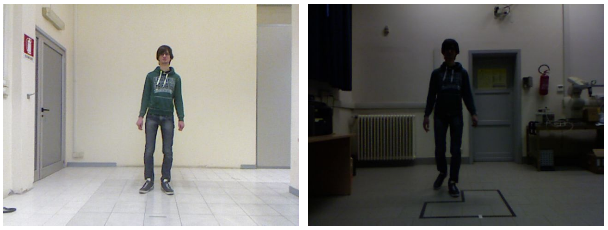
(c) Tracking percentage for every joint

Fig. 4. (a) Re-identification results with respect to the number of recognized joints and (b) number of frames of the BIWI RGBD-ID dataset when varying the number of tracked joints. No frames contain a skeleton with less than 10 tracked joints. In (c), we report the percentage of frames in which each joint is tracked.

exploiting SIFT descriptors.

B. Robustness to illumination changes on the IAS-Lab RGBD-ID dataset

The IAS-Lab RGBD-ID dataset consists of 33 sequences of 11 people acquired using the OpenNI SDK and the NST tracker. For every subject, the *Training* and *TestingB* sequences were collected in different rooms, with strong illumination changes caused by the different auto-exposure level of the Kinect in the two rooms. This kind of issue can be noticed from the sample frames reported in Fig. 5. In Fig. 6, we show the CMCs obtained with the main 2D and 3D descriptors we tested and with the approach we explained in Sec. III-A which computes a global RGB histogram for every person. As it can be noticed, SIFT is again the best descriptor, with a rank-1 very close to 100%. BRIEF and ORB also obtain very good results, directly followed by the PFHRGB descriptor. Unlike in the BIWI RGBD-ID dataset, the color histogram approach obtains poor results on this dataset, probably because of the strong illumination changes, while our texture-based approach maintains high performance.



(a) Training

(b) TestingB

Fig. 5. Sample frames of the same person from *Training* and *TestingB* sets. Strong illumination differences can be noticed.

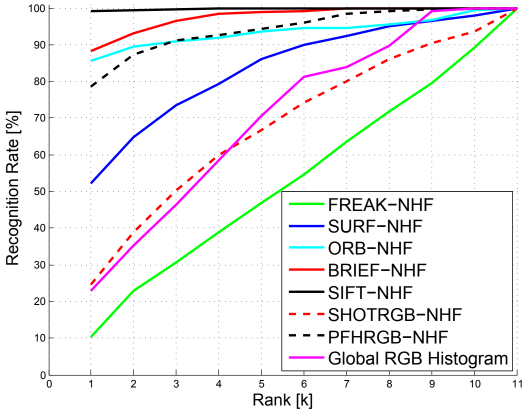


Fig. 6. Re-identification results on the IAS-Lab RGBD-ID dataset.

C. Multi-frame results

In Table I, all re-identification results are reported for all the approaches and datasets we tested in this work. The multi-frame score presented in Sec. II-F is also reported for the BIWI RGBD-ID dataset and the IAS-Lab RGBD-ID dataset. It can be noticed how, if we consider this sequence-wise result, our SPS signature coupled with the SIFT descriptor allows to re-identify all the people of the two tested datasets (R1-Multi = 100%) and in general the re-identification percentage is 10-30% higher than the rank-1 computed for the single-frame re-identification.

D. Further comparisons on the CAVIAR4REID dataset

In order to validate our re-identification approach with respect to the best re-identification approaches in literature, we performed some tests on one of the most challenging datasets which is used for evaluating re-identification in video surveillance scenarios, the CAVIAR4REID dataset. This dataset contains only RGB information, thus we manually annotated the skeleton joints on all the images in order to compute our SPS signature at skeleton keypoints. We distinguished between visible and non-visible joints and we also released our annotations[8] in order to allow further comparisons with our method. Since the training and testing frames are collected from different cameras in this dataset, many joints which are visible in the testing images are not visible in the training images and vice-versa. Therefore our TJ matching approach is not applicable to this dataset and we selected the AJ matching instead, which computes the $\mathbf{SPS}^{AJ}$ signature by concatenating descriptors at all keypoints, regardless if they are tracked or not and substitutes the descriptors for the occluded joints with those computed at their symmetric joints. We performed the same single-frame and multi-frame tests (with M=3 and M=5) described in [2] and we reported the results in Fig. 7. Once again, our SPS signature which exploits the SIFT descriptor obtained the best rank-1 scores for all the tests, thus outperforming both the CPS approach [2] and the SDALF descriptor [1], even though this dataset was targeted to video surveillance applications and the image resolution was very low.

E. Discussion on computational complexity

Table I also reports the times needed for classifying one frame for our SPS signature and the other approaches we tested. For matching with Nearest Neighbor, we exploited KD-Trees and FLANN[9] based matcher, which improve time performance of about one order of magnitude with respect to the brute force algorithm. Our tests were performed with a C++ implementation running on an Intel®Core™i3 CPU M330 @ 2.13 GHz with 4 GB DDR3 RAM.

The SPS signature with 2D descriptors results to be the fastest approach among all the techniques evaluated in this work. In particular, BRIEF is the fastest algorithm and obtains very good re-identification results, even though inferior to SIFT. 3D features are almost one order of magnitude slower than 2D features, thus preventing their use in real time applications. Finally, we show that the algorithm which uses SIFT keypoint detector to select keypoints results to be 2 times slower in the extraction phase and 10 times slower in the matching phase with respect to our skeleton-based approach.

IV. Conclusions

In this paper a novel approach to people re-identification in RGB-D data has been presented. This approach builds on the assumption that very stable keypoints can be detected on human targets by means of a skeletal tracker, and exploited to evaluate signature by means of 2D and 3D feature extractors. This idea was developed considering several features and matching methods and overcoming the instabilities that still affect skeletal trackers.

The novel re-identification system presented in this paper has been extensively tested using both video-surveillance datasets for comparing this novel approach to the state-of-the-art, and exploiting newly created datasets, that are capable of highlighting the great advantages offered by our approach. This re-identification method is particularly suited for robotic applications dealing with humans, since it offers superior performance, it exploits sensors commonly available on most autonomous robots and runs in real-time.

[8] http://robotics.dei.unipd.it/reid.

[9] http://www.cs.ubc.ca/research/flann.

References

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360 –2367, june 2010.

[2] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 2, page 6, 2011.

[3] Kai Jungling and Michael Arens. Feature based person detection beyond the visible spectrum. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2009)*, pages 30–37.

[4] Kyongil Yoon, David Harwood, and Larry Davis. Appearance-based person recognition using color/path-length profile. *Journal of Visual Communication and Image Representation (JVCIR 2006)*, 17(3):605–622.

[5] M Bauml and Rainer Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 291–296. IEEE, 2011.

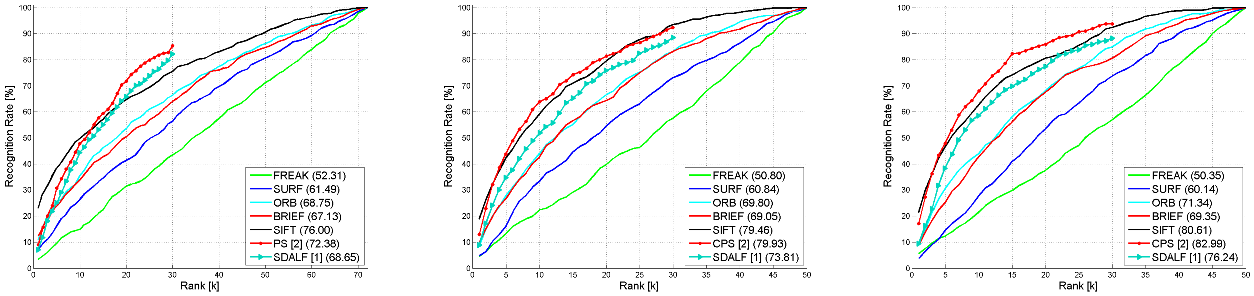
[6] Lei Hu, Shuqiang Jiang, Qingming Huang, and Wen Gao. People re-detection using adaboost with sift and color correlogram. In *Proc. of the 15th International Conference on Image Processing (ICIP 2008)*, pages 1348–1351.

[7] D. Baltieri, R. Vezzani, R. Cucchiara, A. Utasi, C. Benedek, and T. Sziranyi. Multi-view people surveillance using 3d information. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops 2011)*, pages 1817–1824.

[8] J. Oliver, A. Albiol, and A. Albiol. 3d descriptor for people re-identification. In *Proceedings of the 21st IEEE International Conference on Pattern Recognition (ICPR 2012)*, pages 1395–1398.

[9] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *ECCV Workshops 2012*, pages 433–442. Springer.

[10] Jamie Shotton, Andrew W. Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.



(a) Single-frame

(b) Multi-frame (M=3)

(c) Multi-frame (M=5)

Fig. 7. Re-identification results on the CAVIAR4REID dataset. For every approach, nAUC is reported within brackets.

TABLE I

SUMMARY OF RE-IDENTIFICATION ACCURACY AND COMPUTATIONAL TIMES FOR THE MAIN APPROACHES PROPOSED AND COMPARED IN THIS WORK. FOR THE BIWI RGBD-ID DATASET AND THE IAS-LAB RGBD-ID DATASET, THE RESULTS REFER TO TESTS PERFORMED ON FRAMES WITH ALL JOINTS TRACKED. 20 JOINTS ARE USED FOR THE BIWI RGBD-ID DATASET, WHILE THE NHF APPROACH IS REPORTED FOR THE IAS-LAB RGBD-ID DATASET. FOR THE CAVIAR4REID DATASET, TESTS HAVE BEEN PERFORMED ON ALL FRAMES AND WITH ALL SKELETON JOINTS (AJ SIGNATURE).

	Timings (ms)		**BIWI RGBD-ID**			**IAS-Lab RGBD-ID**			**CAVIAR4REID (M=5)**	
Approach	**Extraction**	**Matching**	**Rank-1**	**nAUC**	**R1-Multi**	**Rank-1**	**nAUC**	**R1-Multi**	**Rank-1**	**nAUC**
SIFT	185.45	0.00045	85.2	98.23	100.0	99.2	99.86	100.0	21.4	80.61
SURF	12.45	0.00072	31.0	78.52	75.0	52.3	84.35	72.7	3.8	60.14
BRIEF	9.58	0.00024	62.4	91.16	92.9	88.3	97.68	90.9	8.8	69.35
ORB	12.84	0.00024	53.3	88.12	89.3	85.6	93.83	90.9	9.2	71.34
FREAK	15.64	0.00048	11.1	56.90	28.6	10.2	55.33	9.1	5.6	50.35
PFHRGB-NHF	894.74	0.00137	48.5	87.01	67.9	78.7	94.31	100	-	-
SHOTRGB-NHF	622.17	0.01482	23.0	81.14	39.3	24.6	69.56	45.46	-	-
FPFH-NHF	1317.27	0.00021	12.4	62.18	32.1	-	-	-	-	-
PFH-NHF	765.37	0.00095	11.1	63.54	21.4	-	-	-	-	-
SHOT-NHF	616.09	0.00347	5.4	56.38	10.7	-	-	-	-	-
SIFT-keypoints	413.28	0.0035	63.4	89.99	71.43	-	-	-	-	-
Global RGB Histogram	352.16	0.0075	78.4	98.93	85.7	23.0	71.60	27.3	-	-
Local RGB Histogram	309.15	0.0085	53.4	94.84	75.0	-	-	-	-	-
SDALF [1]	-	-	-	-	-	-	-	-	9.4	76.24
CPS [2]	-	-	-	-	-	-	-	-	17.2	82.99

[11] Morgan Quigley, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibs, Eric Berger, Rob Wheeler, and Andrew Ng. Ros: an open-source robot operating system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[12] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[13] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[14] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV 1999)*, volume 2, pages 1150–1157.

[15] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June.

[16] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: binary robust independent elementary features. In *Proc. of the 2010 European Conference on Computer Vision (ECCV 2010)*, pages 778–792. Springer.

[17] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proc. of the 2011 IEEE International Conference on Computer Vision (ICCV 2011)*, pages 2564–2571.

[18] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, pages 510–517.

[19] R.B. Rusu, N. Blodow, Z.C. Marton, and M. Beetz. Aligning point cloud views using persistent feature histograms. In *Proc. of the 2008 International Conference on Intelligent Robots and Systems (IROS 2008)*, pages 3384–3391.

[20] R.B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Proc. of the 2009 International Conference on Robotics and Automation (ICRA 2009)*, pages 3212–3217.

[21] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *Proc. of the 2010 European Conference on Computer Vision (ECCV 2010)*, pages 356–369. Springer.

[22] Federico Tombari, Samuele Salti, and Luigi Di Stefano. A combined texture-shape descriptor for enhanced 3d feature matching. In *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP 2011)*, pages 809–812. IEEE.

[23] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti. Evaluation of local features for person re-identification in image sequences. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, page in press. IEEE, 2014.

[24] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, volume 5302, pages 262–275. 2008.

[25] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.