A Multi-Viewpoint Feature-based Re-identification System driven by Skeleton Keypoints

Stefano Ghidoni and Matteo Munaro

Department of Information Engineering, University of Padova
via Gradenigo, 6/B – 35131 Padova, Italy

Abstract

Thanks to the increasing popularity of 3D sensors, robotic vision has experienced huge improvements in a wide range of applications and systems in the last years. Besides the many benefits, this migration caused some incompatibilities with those systems that cannot be based on range sensors, like intelligent video surveillance systems, since the two kinds of sensor data lead to different representations of people and objects. This work goes in the direction of bridging the gap, and presents a novel re-identification system that takes advantage of multiple video flows in order to enhance the performance of a skeletal tracking algorithm, which is in turn exploited for driving the re-identification. A new, geometry-based method for joining together the detections provided by the skeletal tracker from multiple video flows is introduced, which is capable of dealing with many people in the scene, coping with the errors introduced in each view by the skeletal tracker. Such method has a high degree of generality, and can be applied to any kind of body pose estimation algorithm. The system was tested on a public dataset for video surveillance applications, demonstrating the improvements achieved by the multi-viewpoint approach in the accuracy of both body pose estimation and re-identification. The proposed approach was also compared with a skeletal tracking system working on 3D data: the comparison assessed the good performance level of the multi-viewpoint approach. This means that the lack of the rich information provided by 3D sensors can be compensated by the availability of more than one viewpoint.

Keywords: People re-identification, people tracking, body pose estimation, camera networks, multi-view skeletal tracker.

1. Introduction

People re-identification is a topic that has been widely addressed in the literature, as it is a crucial capability in several fields, including intelligent video surveillance, service robotics, assistive robotics and many others. While people re-identification is a widely addressed topic in the field of image processing, techniques based on the analysis of 3D data became popular only recently. This was caused by the introduction of low-cost, high-resolution RGB-D (RGB-Depth) sensors into the market, which pushed mobile robotics towards 3D vision and gave strong impulse to the development of the Point Cloud Library (PCL)[1] [1]. First released in 2011, four years later it is considered the *de facto* standard for the processing, filtering and storage of 3D data, with the plus of being open source, and it is used throughout the world.

The introduction of RGB-D data had a strong impact on many applications connected to mobile robotics. As an example, consider people tracking systems: moving from RGB to RGB-D sensing provides, on one hand, enhanced accuracy in the metric measurements (e.g. target location and dimensions). On the other hand, this offers new cues that can be observed and exploited for tracking purposes, like the body shape. This also had a positive impact on the algorithms for skeletal tracking (i.e., the detection of the body pose): those based on 3D data provide reliable results, while the 2D counterparts still show some difficulties, even though great improvements have been achieved [2].

The picture outlined so far suggests that 3D sensing is the way to move robot vision one step forward; however, there are some applications and environments that are more suitable for 2D vision. The most important limiting factor is the range of the mentioned low-cost 3D sensors, that can be (in the case of the

[1] Available at: `www.pointclouds.org`

Kinect 2) up to 8-9 m in the best working conditions [3]. This means that robot navigation and people and object detection cannot be based on 3D sensing in large industrial environments, or outdoors (where the working range is reduced) unless more sophisticated and expensive sensors are employed, as it is the case of the Velodyne sensor [4]. Likewise, intelligent video surveillance applications usually deal with events that occur at a rather long distance, and are rarely based on RGB-D sensors.

In summary, the availability of low-cost RGB-D sensors created two different research lines that tackle similar problems based on different sensory information. While this gave new impulse to the research activity and generally increased the performance level of robotic perception, an important drawback was also introduced: a more difficult interoperability among systems based on different sensors. As an example, consider a service robot moving in an environment where an intelligent video surveillance system making use of a camera network is also operating. If the robot is running a people tracking algorithm based on RGB-D data, it will make use of models that rely on 3D information: thus, a comparison between its tracks and those generated by the video surveillance system is difficult to perform.

This paper goes in the direction of filling the gap: a novel re-identification system for camera networks is presented, based on the analysis of RGB data only, that is, 2D images, without relying on 3D information. The lack of depth data is counterbalanced by the observation of the scene from multiple viewpoints, that provides additional information, exploited to improve the performance of the body pose estimation. The comparison between the multi-viewpoint 2D approach and an open source single-viewpoint 3D skeletal tracker, presented in Section 4, demonstrates that relying on 3D data is not always the best choice.

The system presented in this paper aims at generating a body model suitable for re-identification which is similar to the model developed in [5]; however, given the difference in the input data (multiple viewpoint 2D data in this case, single viewpoint 3D data in [5]), the processing needed to achieve such similar

representation is different. In particular, since the skeletal tracker exploited in our previous system needs RGB-D data, it is not possible to exploit the same algorithm using RGB data only. Furthermore, dealing with multiple viewpoints, an additional challenge needs to be faced: the association of the same person among the different views analyzed. This means that a novel processing pipeline should be developed, with the constraint of providing, for each person in the scene, a signature for re-identification which can be compared with those evaluated from a single RGB-D sensor: in other words, even though the multiple view RGB and the single view RGB-D systems are based on different algorithms, they need to provide the same type of output data, which is crucial in order to compare observations performed using different sensory systems.

Overall, the system proposed in this paper presents the following elements of novelty: i) a method for merging together estimations of the body pose made from different viewpoints: multiple hypotheses for each viewpoint are considered, and the final merging phase is performed in the 3D space; ii) the application of a matching technique that is capable of associating the correct body pose among different views when multiple people are found in the scene; iii) the application of the re-identification method already developed for 3D data in this new context; iv) extensive tests on a publicly available dataset for re-identification applications.

It is important to observe that the method for merging together multiple views is decoupled from the single-view pose estimation algorithm, in contrast to what is proposed in [6]. This second option is focused on boosting performance, since the final 3D location of each joint is evaluated directly from all the images. On the other hand, our approach offers the advantage of being modular (because any 2D single-view body pose estimation algorithm can be used), simpler and with a very low computational cost, as illustrated in Section 4.5. Even though our approach does not have access to the source image while determining the 3D locations of the image joints, it considers several possible body poses guessed by the skeletal tracker, each one "voting" for a specific 3D body configuration. The RGB body pose estimation algorithm employed in this work is the Part-Based

Detector (PBD) proposed in [2].

Regarding re-identification, the availability of multiple video sources lets the system observe each skeleton keypoint from several angles. This aspect will be considered in the future developments of the system, because the availability of multiple views of the same keypoint can enrich the target description, e.g., the front and rear part of a person can be seen at the same time by two cameras facing each other. The same advantage is not available in the single RGB-D sensor system described in [5].

The paper is organized as follows: in Section 2, previous work related to skeletal tracking and re-identification is discussed; in Section 3, the multi-view approach to skeletal tracking is detailed, together with the re-identification system exploited for performing the experiments. Experimental results are reported in Section 4, while the concluding remarks are summarized in Section 5.

2. Related Work

People re-identification in images is addressed by observing three main characteristics: color, texture and shape, either considered separately or mixed together. A comparison and evaluation of the most important works in the literature is reported in [7]. The methods which exploit global histograms in RGB or HSV space assume that people can be distinguished by looking at their main colors and they keep the same appearance from every point of view. One of the best color-based approaches in the literature divides the body of each target into smaller parts and evaluates multiple histograms, one for each part [7, 8]. This method is simple and effective, but suffers from two main flaws: it fails when the same target is seen in different illumination conditions, 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 and exploit descriptors evaluated on a set of keypoints to generate the signature of a target. Performance is therefore strongly related to the characteristics of the set of descriptors selected, including the capability of the key-

point detector to select stable features. This approach is widely used in the literature [9, 10, 11] 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 was also used together with histograms [12]. However, approaches based on local appearance feature extraction are usually computationally heavy because many features have to be matched at every frame and many mismatches can occur.

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. Moreover, efficient skeleton tracking algorithms [13] have been released which provide 3D position and orientation of the body joints from 3D data; some works also exploit temporal information [14] by tracking the skeleton joints. 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 can 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 [15, 16, 17, 18, 19]. In [15] authors propose to build a body model of each person by projecting texture from different views to a mummy-like 3D model, which is then used for merging together short-term tracks that are recognized to belong to the same person. The matching stage aims at finding correspondences between pairs of models using shape and color information in order to calculate their similarity.

To overcome the problem of keypoint detection and matching, we recently proposed an approach for fast and accurate re-identification based on RGB-D data [5] that exploits the body joints, extracted by means of a skeletal tracker, by taking them as keypoints on which features (e.g. SIFT) are evaluated. The high stability of such keypoints leads to a very good re-identification performance.

Given the effectiveness of skeleton keypoints for people re-identification, it

would be interesting to migrate such approach from 3D sensors to camera networks, where the skeleton can be robustly estimated by exploiting video flows acquired from multiple viewpoints. Some approaches to tackle this problem already exist: [6] proposes a multi-view pictorial structure algorithm leading to a 3D model of the person. The approach was further developed very recently [20], and other approaches were also recently discussed in the literature [21], however, the main problem affecting 2D pose estimation algorithms is that the performance strongly depends on the viewpoint: when a person is seen from a lateral point of view, the self-occlusions become very strong and jeopardize the detection; this can be solved only by using more than one point of view.

3. System Structure

The re-identification system proposed in this paper is meant to take advantage of the multiple video sources available in camera networks. The first requirement on which it is designed is that the camera network is fully calibrated (the intrinsic and extrinsic parameters are available to the system). In case the network includes active cameras, information about the change in pose and zoom needs to be delivered at the same frequency at which images are acquired. The second requirement involves the way cameras are placed and oriented: since the re-identification system improves the results of RGB skeletal trackers by considering more than one view of each target, the cameras need to frame overlapping regions from different viewpoints.

At a high level, the proposed system acquires 2D data, processes it in the 3D world coordinate domain, and goes back to the 2D image domain for the final processing. The main steps to get a target signature are as follows:

- **2D** run the skeletal tracker on each input image (see Sections 3.1 and 3.2);
- **2D→3D** project the keypoints into the 3D world coordinate system (see Section 3.3);

- **3D** merge the results from different viewpoints, evaluating a single body pose in the 3D domain for each target (discussed in Sections 3.3 and 3.4);
- **3D→2D** reproject the final body pose to each view (detailed in Section 3.5);
- **2D** evaluate the feature vectors on the reprojected joints (as discussed in [5]).

It is important to recall that the 3D information is recovered from the 2D images analyzed and the calibration data: no depth data is made available by the sensors employed.

The final signature of each target is then obtained by concatenating all the feature vectors related to the body joints in an ordered way, as detailed in [5] — this signature is called Skeleton-based Person Signature (SPS). The final association among the different candidates (the re-identification activity properly said) is performed by comparing the SPS of all targets using a minimum distance criterion.

3.1. Single-view Skeletal Tracking System

The starting point of the multi-view skeletal tracker is an algorithm for body pose estimation on RGB images. As discussed in Section 2, the evaluation of the body pose in 2D images can still be considered an open problem, while the performance level reached by systems working on RGB-D data is surprisingly high. The system chosen for analyzing the input image is the Part-Based Detector (PBD) [2], which models deformable objects (including the human body) using smaller, non-oriented parts (hence the name). The algorithm analyzes local appearance, geometrical relations between parts, and co-occurrence relations between part mixtures: this leads to a high-level representation of the object. The PBD was chosen because it offers two key advantages over other approaches: a state-of-the-art performance level and an efficient C++

open-source implementation[2], which is very well suited for being reused inside another project.

The PBD is capable of detecting several different deformable objects, depending on the model loaded. A pre-trained human model, provided by the authors, was used in this work[3]: it is made of 26 parts, that represent the whole body; the largest ones (e.g. arms and legs) are decomposed into smaller parts, as the PBD works best relying on small, unoriented parts.

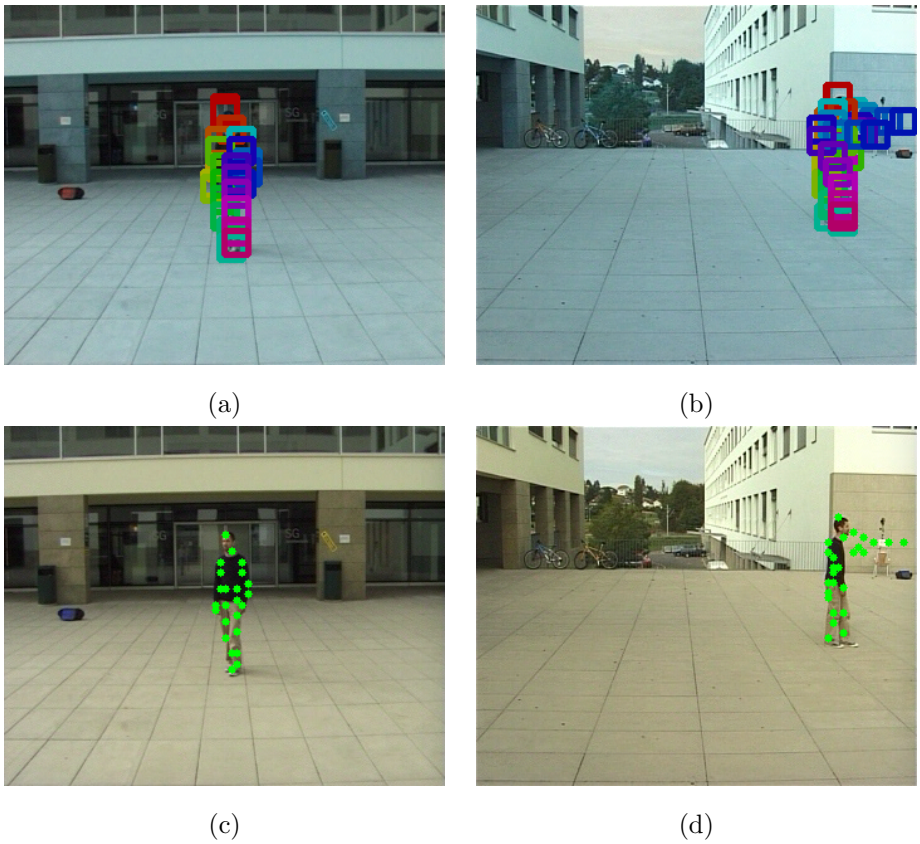
From the geometrical point of view, the output provided by the PBD is a set of rectangles (parts) in image coordinates, as it can be seen in figure 1 (a-b), which show the results for a pair of images acquired by a camera network. Such rectangles need to be transformed into a set of body joints, as these are more suitable for the triangulation tasks performed by the multi-view stages of our system. The conversion to body joints is performed by substituting each rectangle with the crossing point of its two diagonals. This is the only meaningful choice that does not require the analysis of the pixel content inside each rectangle: recalling that the PBD approach does not measure the body part orientation, each rectangle contains a piece of the body that might be oriented in any direction; for example, a portion of an arm can be placed along a horizontal, vertical or diagonal line, and this information is not provided. The center of each rectangle is therefore the only area that is surely placed on the foreground, which is very important for the re-identification part of the system. The results of the conversion of the PBD output in figure 1 (a-b) is reported in (c-d).

3.2. Candidate Clustering

The PBD creates an extremely low number of false detections (meant as detections in areas of the image that do not contain any person), while it usually generates a very large number of candidates, in the order of hundreds, for each

[2]The source code is available at https://github.com/wg-perception/PartsBasedDetector

[3]The "PARSE_model.mat" model was used, which is available at http://www.ics.uci.edu/~dramanan/software/pose/.



(a) (b)

(c) (d)

Figure 1: Sample of the output of the Part Based Detector on two images taken from two cameras with overlapping fields of view (a-b) and the corresponding keypoints evaluated by the system.

person framed in the input image. The number of candidates decreases as the size of the person in the image becomes smaller, and reaches the working limit of the PBD. Among the many candidates generated, some of them have a correct body pose, while others show some errors, mainly affecting arms and legs. This was experimentally found to be weakly correlated with the confidence value which is provided together with every detection. However, it was observed that several correct detections occur among the candidates with highest confidence.

As important information can be inferred by analyzing how candidates group around the people found in the input image, a clustering phase was introduced, in order to organize the candidates provided by the PBD into groups, called detections. The goal of clustering is to organize detections into groups, each one referred to a single target found in the image. A set of candidates can be considered as a set of different guesses about the body pose a target. The number of candidates in a group is related with the confidence of a detection, because a larger number indicates a stronger detection.

Given a pair of candidates generated by the PBD, it is possible to verify whether they refer to the same person in the scene by checking their overlap in the image. It is meaningful to perform this analysis in the 2D image coordinates, since candidates are referred to this reference system, based on 2D information.

The order in which candidates are provided by the PBD can be considered random. This means that, if more than one person is present in the input image, two consecutive candidates can refer to different persons, and switches among targets can occur in any order. Starting from the highest confidence value, candidates are grouped into clusters that are composed by the elements that overlap on the same region of the image. The superimposition is calculated evaluating the bounding box including all the skeleton keypoints, and imposing a maximum distance constraint among the central points of such bounding boxes.

Since the PBD generates a lot of candidates, there is a limit N on the number of elements included in each cluster (this was set to 5 for performing the experiments). This means that, for each person found in the scene, the best N guesses provided by the PBD are considered: this way the best guess is

included in the set, even though it does not have the highest confidence value. The value of N should be carefully chosen in order to reduce the noise introduced by wrong candidates, that are the ones with lower confidence values.

3.3. Candidate Triangulation and Filtering

Once clusters of candidates are calculated in all the input images provided by the camera network, the multi-viewpoint processing is triggered. All pairs of views are considered: given a pair (V_a, V_b), all the n clusters belonging to the first view, $\{C_0^a, \ldots, C_{n-1}^a\}$ are compared against the m clusters found in the second view, $\{C_0^b, \ldots, C_{m-1}^b\}$. Each cluster-to-cluster comparison is performed by projecting the joints to the 3D world coordinate system: as it is well known [22], each joint in the image domain generates a 3D line. Since each cluster is composed by N candidates, each one being composed by 26 joints, a total of $26 \times N$ lines is generated from each cluster from each view. It should be pointed out that our clustering algorithm selects *no more than* N candidates for each cluster, but they can also be less than N, as the PBD may provide a reduced number of candidates for a given target. However, this happens only when the detection is hard to perform: in the general case, each cluster includes the maximum number of candidates. Without loss of generality, it will be assumed that each cluster is composed of N candidates.

It should be noted that, even though the cameras are intrinsically and extrinsically calibrated, it is not possible to associate clusters in two different views based only on their location on the images, as this would require knowledge about the position in the 3D world domain (which is the information that is being evaluated). Each of the cluster couples must then be considered as a possible match.

A couple of clusters C_i^a and C_j^b, belonging to views V_a and V_b, respectively, is matched by considering each joint type separately: for each type (say, the left hip), N 3D lines are available from each view. All the intersection points are then evaluated, generating N^2 3D points, $\{p_0^{i,j}, \ldots, p_{N^2-1}^{i,j}\}$. Strictly speaking, intersection points can almost never be found, as the 3D lines considered are

generally skewed; each intersection is therefore approximated by the center point of the segment representing the minimum distance between the given line pair. The value of the minimum distance is taken into account to evaluate the quality of the intersection, as it will be detailed in Section 3.4.

The intersection points represent a set of hypotheses for the location of each body joint in world coordinates. Since they are a set of points in 3D, in fact they form a point cloud, and might be processed accordingly. The final goal of such processing is the identification of a single 3D point that will be considered the real position of the body joint in world coordinates. In figure 2 it is possible to see the result of a triangulation for a target, in perspective view (a) and from a bird's eye view (b).

The point cloud is expected to include a subset of points that are rather close to each other, generated by the combinaton of right guesses from both views, plus some more points that are in completely different locations, generated by the triangulations involving wrong poses from at least one of the views. A statistical outlier removal algorithm is employed to filter out such cases. For each point, the filter measures the distances to all the others, and evaluates the mean and variance of their distribution (assumed to be gaussian): the point is then discarded if such values are outside the acceptance range[4]. Once the outliers are removed, the cloud centroid is evaluated and assumed to be the right location of the given keypoint. After this procedure is repeated for all the 26 keypoints of the skeleton, a set of 26 joints in 3D coordinates is available, which is assumed to be the body pose in 3D.

3.4. Association of Triangulated Clusters

Given two views, all the possible combinations of their clusters are analyzed. The cluster triangulation is applied to *every* couple of clusters for each given couple of views. Consider, as an example, the case of two persons P_0 and P_1

[4]The algorithm employed is the one implemented in the StatisticalOutlierRemoval class found in PCL. The parameters used are $StddevMulThresh = 1.0$ and $K = 5$.

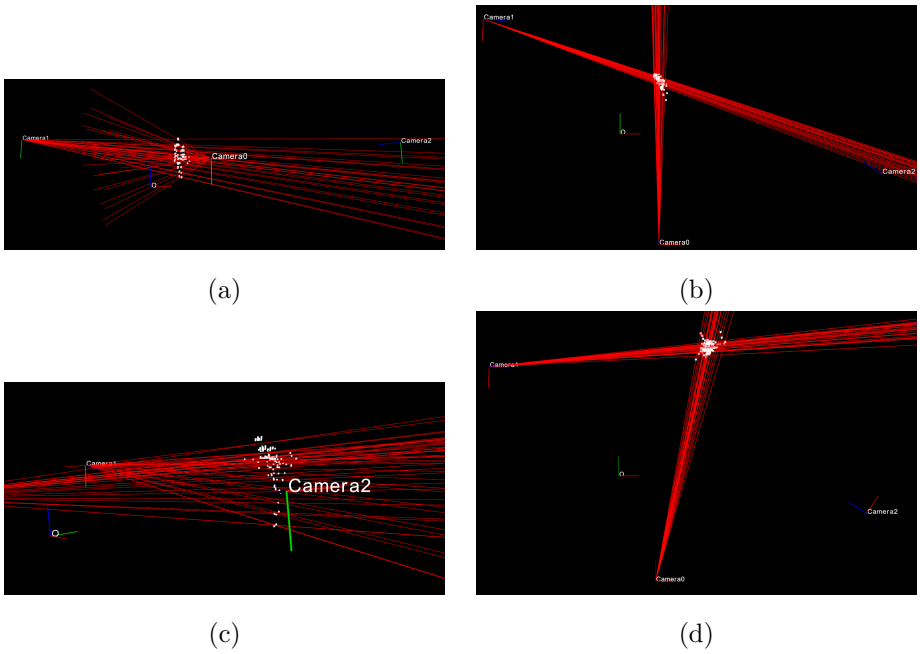


Figure 2: Triangulations of a target: the scheme shows the perspective view (a) and the bird's eye view (b) when a correct match occurs. The camera network includes three cameras, and the world reference system (O) was also added. The red lines are the projections of the keypoint found on the images into the 3D world; the white dots represent the intersection points. A human shape is recognizable in (a). In (c) and (d) the same scheme is reported for a wrong match: it is possible to see that the points are not roughly vertically aligned, but rather they are placed along a curve (c), and the projections of the the points onto the ground cover a rather large area (d).

that are in the field of view of the system, and are framed at the same time from the two views V_a and V_b. The first person will then generate two projections, P_0^a and P_0^b, and the same will happen for the second person, that will generate P_1^a and P_1^b. The system will perform the following comparisons:

- **Match 1:** P_0^a against P_0^b;
- **Match 2:** P_0^a against P_1^b;
- **Match 3:** P_1^a against P_0^b;
- **Match 4:** P_1^a against P_1^b.

Matches 1 and 4 will provide reasonable results, while 2 and 3 will lead to point clouds that might show artifacts and other distorting effects, depending on how the projections of such unrelated points intersect. An example of the distorting effects can be seen in figure 2 (c): the intersection points are not placed around a vertical axis, but rather, they follow a curved line; also when seen from a top view (d), the points cover a rather large area.

In the real working cases, when the correct matches are not known, this is useful to understand whether the two matched cluster refer to the same person or not. For this purpose, we propose a distance function that measures: i) the distance between head and left hip in the 3D domain, D_{HH}; ii) the distance between the head and the centroid of the two feet, projected onto the ground, D_{HF}; iii) the average of the minimum distance between the lines that generate the "intersection" points (again, recall that the lines are in fact skewed, thus such distance is not zero), D_L.

Based on the experimental observations performed, a target value was defined for each of the mentioned distances, which is 1 m for D_{HH} and 0 for D_{HF} and D_L. The value chosen for D_{HH} is quite high because the PBD finds the hips close to the top part of the thighs. The distance function for measuring the quality of a given association is therefore:

$$D = \begin{cases} \infty & \text{if} D_{HH} > MAX_{HH} \\ \frac{1}{3}\left(|D_{HH} - 1.0| + \frac{D_{HF}}{w_{HF}} + \frac{D_L}{w_L}\right) & \text{otherwise} \end{cases}, \quad (1)$$

	$\mathbf{c_1^A}$	$\mathbf{c_2^A}$	$\mathbf{c_3^A}$
$\mathbf{c_1^B}$	d_{11}	d_{12}	d_{13}
$\mathbf{c_2^B}$	d_{21}	d_{22}	d_{23}

(a)

	$\mathbf{c_1^A}$	$\mathbf{c_2^A}$	$\mathbf{c_3^A}$
$\mathbf{c_1^B}$	∞	0.8	0.6
$\mathbf{c_2^B}$	1.5	∞	0.5

(b)

	$\mathbf{c_1^A}$	$\mathbf{c_2^A}$	$\mathbf{c_3^A}$
$\mathbf{c_1^B}$	0	1	0
$\mathbf{c_2^B}$	0	0	1

(c)

Figure 3: Example of application of the Munkres algorithm for cluster matching between two images A and B where three and two clusters are found, respectively. At first, the distance matrix (a-b) is computed by using equation 1, then the Munkres algorithm solves the association problem and provides a matrix (c) in which each 1 corresponds to a couple row-column that identifies the best matches. In this case, the matches are (c_1^B, c_2^A) and (c_2^B, c_3^A). Cluster 1 in image A (c_1^A) has no correspondence in image B, thus it is discarded.

where the values chosen for the threshold and weighting factors are $MAX_{HH} = 1.3$, $w_{HF} = 0.3$ and $w_L = 0.2$. The choice of placing a threshold on D_{HH} comes from the observation that values outside that range correspond to wrong clusters in all cases.

The distance in (1) is evaluated for all the triangulations given by the possible combinations of clusters in the two views considered. The final association among the clusters is given by selecting the matches with shorter distance: this is modeled as a Global Nearest Neighbor (GNN) problem, which is then solved using the Munkres algorithm[5], described in [23, 24].

Figure 3 reports an example of cluster association when two images A and B contain a different number of clusters, three and two, respectively. The combinatorial optimization algorithm we use allows to obtain the globally best associations in polynomial time. The best triangulations selected are considered to belong to the real persons that are present in the framed scene. If a cluster in one image is not associated to a cluster in the other image, it is discarded without further processing.

It should be noted that the values of D_{HH}, D_{HF} and D_L are taken into account for associating the clusters found in the different views. Once this is done, each couple of associated clusters is considered, and all the joints of the

[5]http://csclab.murraystate.edu/bob.pilgrim/445/munkres.html

same type (e.g., all the joints referring to the neck) are triangulated using the technique described in Section 3.3, leading to a single 3D location for that type of joint. This is then repeated for all the joint types of the skeleton.

3.5. Reprojection of Triangulated Clusters

The body pose in world coordinates shall be used to drive the evaluation of the features that will in turn compose the SPS. The 3D points are therefore projected onto the image planes of the cameras composing the network, which is possible thanks to the accurate calibration of intrinsic and extrinsic parameters of the sensors.

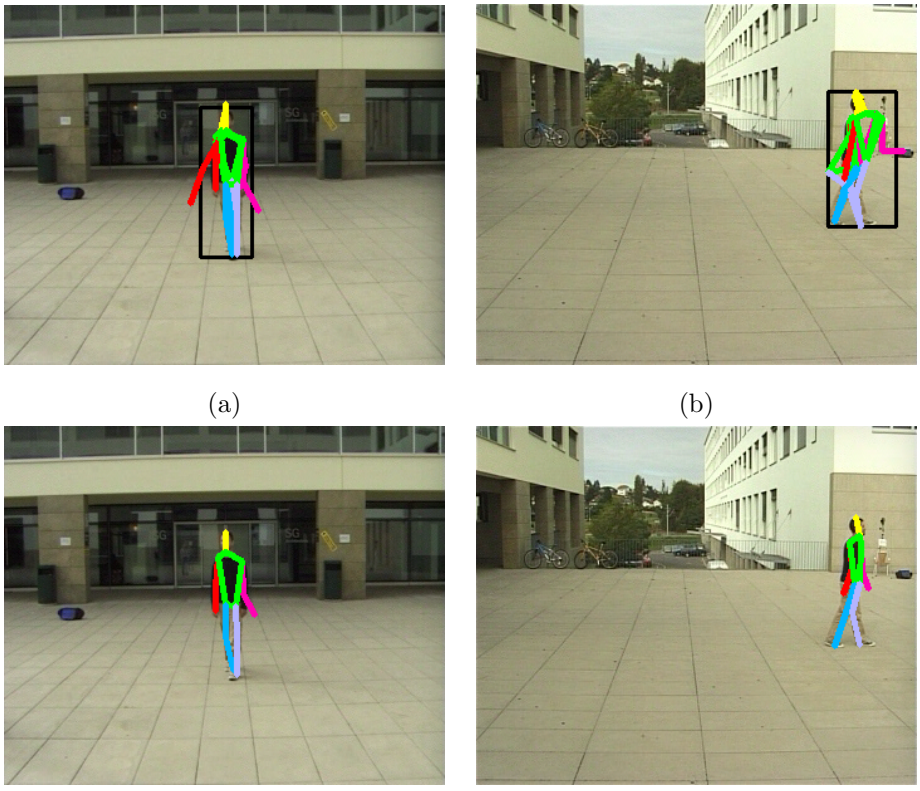
Figure 4 (a-b) shows the results of the clustering from two views of the system: some errors in the evaluation of the body pose, especially about the arms, can be seen. In (c-d), on the same frames are reported the results after the triangulation and association phases, which is the final output of the system. The advantages provided by the multi-view processing can be recognized.

After the reprojection, the re-identification system can proceed in the same way as proposed in [5]; in particular, it was demonstrated that the SPS shows the best performance when used in combination with 2D features evaluated in the RGB image domain: this means that a system based on a camera network does not lose performance if compared with systems equipped with RGB-D sensors.

4. Experiments

This section presents the results obtained by the proposed approach, both in terms of accuracy of the detected body pose, and of re-identification performance. In paragraph 4.2, the multi-viewpoint 2D skeletal tracker is compared with a different approach, based on single-viewpoint 3D data, which is a sensory data widely used since the introduction of inexpensive 3D sensors in the last years.

We tested the multi-view skeletal tracking algorithm described in this paper on the first sequence of the *Campus* dataset, a public dataset proposed in [25].



(a) (b)

(c) (d)

Figure 4: Two views showing the clustered skeletons (a-b) and the reprojections of the 3D triangulated points (c-d). It is possible to see that the processing in 3D world coordinates rejects the outliers that cause the wrong estimations of the arms.

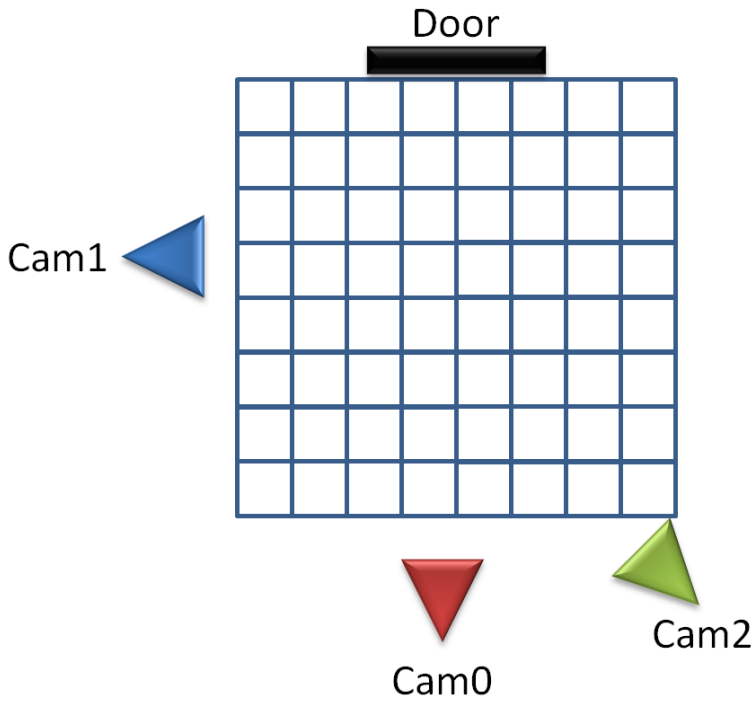


Figure 5: Placement of the cameras for the *Campus* dataset.

This sequence contains three views of the EPFL campus with the cameras placed as outlined in Figure 5. Sample images from the three cameras are reported in Figure 6; up to four people are simultaneously walking in front of them and three out of these are annotated on the ground truth provided in [26].

4.1. Test on skeleton accuracy

To compare the accuracy of our multi-view skeleton estimation methods with that of a single-view approach, we computed the distance of the estimated skeleton joints to the ground truth provided in [26]. Since the ground truth was composed of only 14 joints, we selected the same joints from the output of the algorithms to test. Then, we computed the skeleton-to-skeleton distance over 440 frames, that are the frames for which the output of all the algorithms and the ground truth are available. In Table 1, we reported the average joint

(a) Cam0 (b) Cam1 (c) Cam2

Figure 6: Sample images showing the three points of view of the *Campus* dataset.

Table 1: Pixel accuracy of skeleton estimation on the *Campus* dataset compared against the ground truth provided in [26].

Single-view	**BM Multi-view**	**AM Multi-view**
18.2	15.7	15.1

distance to the ground truth for the joints estimated with three different skeletal trackers:

- *single-view*: this is the part-based body detector proposed in [2], that works from a single image;
- *best match (BM) multi-view*: this is the multi-view approach described in Section 3 that keeps only the skeleton with highest confidence among the skeletons of every cluster;
- *all matches (AM) multi-view*: this is the multi-view approach described in Section 3 that combines all the skeletons in every cluster.

It can be noticed how the multi-view approaches allow to obtain a 20% improvement with respect to the method exploiting a single view, thus validating the effectiveness of the proposed algorithms. As expected, the multi-view algorithm exploiting all matches (*AM multi-view*) turned out to be more accurate than the one only exploiting the best match (*BM multi-view*).

4.2. Comparison with a 3D skeletal tracker

We compared the multi-viewpoint skeletal tracker previously described with a different approach, based on 3D data. We chose the skeletal tracking system available in PCL [27] with the improvements developed in a project sponsored by Google Summer of Code[6]; such skeletal tracker is based on the people detection system described in [28]. This system was chosen because it is open source[7] like the PBD, and can therefore be tested on any sequence.

The 3D skeletal tracker could not be tested on the *Campus* dataset, which does not contain 3D data; we therefore acquired an additional test sequence, called *3D reference* dataset using three 3D sensors framing one person walking and turning; both 3D and 2D data were recorded from each sensor. No interference among the sensors were observed in the depth images, because the difference among the sensor orientations were enough to let them frame different parts of the subject. Such sequence was acquired indoors in order to avoid artifacts in the 3D data, that can show up when framing objects under direct sunlight. The sequence acquired shows a person waking quite slowly, therefore the experiments were performed analyzing one image out of 10, in order to process body poses that are substantially different from each other; a total of 60 frames (including both 2D and 3D data) was considered.

The sensor placement is similar to the sketch in Figure 5 and the camera poses have been automatically calibrated with the method proposed in [29]. The orientations are the same, but the cameras were placed at a closer distance among each other, in order to frame the subject from a shorter distance, thus reducing the noise affecting the 3D data. Some results are shown in Figure 7 for the 2D approach (a,c) and the 3D approach (b,d). It is possible to see both algorithms providing a correct detection (a,b) and a wrong detection (c,d); however, while errors in the 2D approach usually involve only a portion of the skeleton, failures in the 3D approach often cause the whole skeleton to be

[6]Further details at: http://www.pointclouds.org/blog/gsoc14/aroitberg/index.php

[7]Downloadable from: https://github.com/AlinaRoitberg/pcl/tree/master/gpu/people

Table 2: Accuracy of the skeleton joint detection on the *3D reference* dataset for the single-view PBD (evaluated on the central camera), BM multi-view, AM multi-view and 3D skeletal tracker (improved version of PCL), measured in terms of average distance between detected joint and ground truth.

Single-view	**BM Multi-view**	**AM Multi-view**	**3D PCL**
23.5	22.6	20.8	60.0

wrongly placed (d).

The accuracy of the body pose detection was also measured in terms of the distance between the detected joints and the ground truth, leading to the results summarized in Table 2, which has the same structure of Table 1. Quantitative measurements show that the multi-viewpoint approach guarantees a much higher accuracy, and confirm the qualitative observations derived from Figure 7. The average pixel error for Single-view and BM/AM Multi-view in Table 2 are higher than the corresponding values of Table 1 due to the higher resolution of the images in the *3D reference* dataset with respect to the *Campus* dataset, causing the framed persons to be composed of a larger number of pixels in the former case.

It should also be noted that even the single-viewpoint 2D approach outperforms the 3D approach chosen for the comparison. On one hand, this depends on the choice of the skeletal tracking algorithm chosen for the comparison, and other approaches would have probably shown better performance. On the other hand, this comparison is meaningful, because we selected two approaches that are open source, that can be downloaded from the internet and applied to any dataset — while other systems cannot be employed on any dataset. This leads us to the observation that 3D data can sensibly improve the performance of body pose estimation, but also offers a high level of complexity that can reduce the performance with respect to image-based systems.

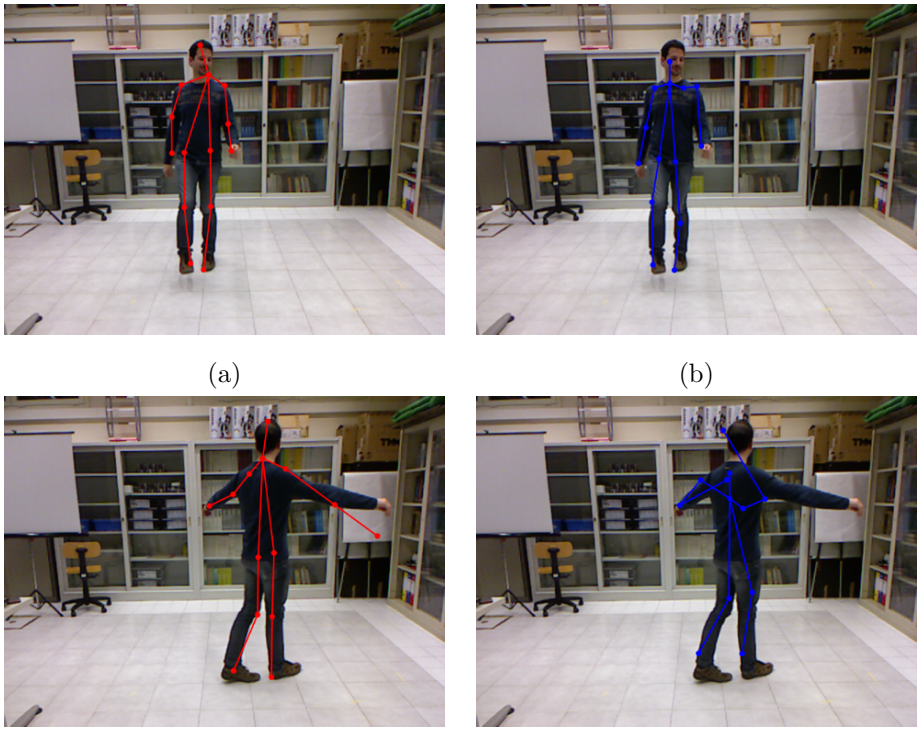


Figure 7: Comparison between the results of the multi-view 2D skeletal tracker proposed in this paper (a) and (c), and the alternative single-view, 3D method (b) and (d). In the first row, examples of good behavior of both approaches are reported; in the second row, examples of wrongly detected skeletons are shown. From (c) it is possible to observe that errors in joint localization affect one limb, but the remaining part of the skeleton is reasonably detected; conversely, errors affecting the 3D skeletal tracker involve the general structure of the body pose. The examples in (c) and (d) are representative of the whole test sequence.

4.3. Test on person re-identification

In [5], we proposed the *Skeleton-based Person Signature* (SPS), a signature for person re-identification based on skeleton information. The accuracy of the skeletal tracker is fundamental to guarantee the stability of the detection of the skeleton keypoints and thus good computation of this signature. In this section, we want to analyze the improvements that can be obtained on person re-identification with the use of the multi-view skeletal tracker described in Section 3.1. In particular, we considered two camera views of the *Campus* dataset (*Cam0* and *Cam1*) as input for creating the multi-view skeletons and we performed two different tests. The restriction to two views only comes from the fact that the Munkres algorithm is able to match elements taken from two sets (represented as the rows and columns of the matrices described in Figure 3).

4.3.1. Test A: Re-identification in the same views

In this test, we applied the re-identification method described in [5] by taking the first 500 frames (one quarter of the whole sequence) as training set and the remaining frames of the same views as testing set. The aim of this test was to evaluate the ability of this person re-identification approach to detect people seen in the past from the same camera, but without the assumption that exactly the same portion of the body is seen in the two occurrences. We compared the re-identification results in terms of rank-1 recognition rate (*rank-1*) and normalized area under curve ($nAUC$) [5] by exploiting the three skeletal trackers we also used for the accuracy test. As ground truth, we used the one provided in [26] and, for comparing the techniques described above on the same input data, we kept only those person instances for which both the part-based people detector and the multi-view skeletal tracking provided a valid skeleton.

As reported in Table 3 and in Figure 8, the *AM multi-view* approach obtained the best re-identification scores (rank-1: 76.4%, nAUC: 86.8%), thus improving the rank-1 recognition rate obtained with the single-view skeletal tracker of about 4%. The *BM multi-view* approach, instead, obtained scores similar or slightly worse with respect to those of the single-view technique, thus suggesting

Table 3: Re-identification scores (*rank-1* and *nAUC*) obtained by exploiting the single-view skeletal tracker and the two versions of the proposed multi-view skeletal tracker for the two tests we performed on the *Campus* dataset.

	Single-view		**BM Multi-view**		**AM Multi-view**	
	Rank-1	**nAUC**	**Rank-1**	**nAUC**	**Rank-1**	**nAUC**
Test A	72.6%	85.5%	71.7%	85.2%	76.4%	86.8%
Test B	65.5%	87.2%	67.6%	86.8%	73.4%	89.5%

that using just one skeleton per cluster for composing the multi-view skeleton is too subject to noise.

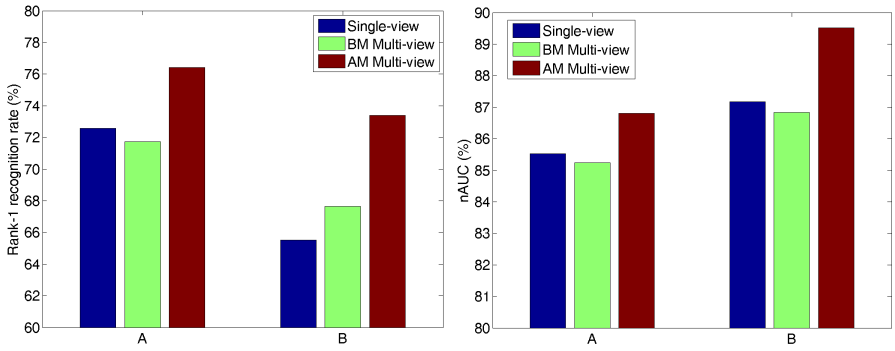
4.4. Test B: Re-identification in a different view

In a network of cameras, two or more cameras can improve the skeleton they use for re-identification with a multi-view approach, but they can also send their multi-view skeleton estimation to further cameras, that can then avoid the computation necessary to perform skeleton estimation or that can exploit a good skeleton even when accurate skeleton estimation would not be possible from their point of view, for example because the person is too far away or in a side pose.

For this reason, we performed a test where the training set was again composed of the first 500 frames from *Cam0* and *Cam1*, but the testing set was composed by all the frames of a third camera (*Cam2*), that was neither used for creating the multi-view skeleton, nor for extracting skeletons for the training set. In particular, we projected the multi-view skeletons estimated by *Cam0* and *Cam1* to *Cam2* image plane thanks to the knowledge of the camera poses. Even in this test, the AM multi-view approach obtained the best recognition results, improving the single-view method of about 8% and the BM multi-view of 6%.

4.5. Computational load

The multi-view skeletal tracker works on a small amount of data, namely the set of joints of each person seen in the scene (considering a number of different



(a) Rank-1 (b) nAUC

Figure 8: Illustration of the re-identification scores (*rank-1* and *nAUC*) obtained by exploiting the single-view skeletal tracker and the two versions of the proposed multi-view skeletal tracker for the two tests we performed on the *Campus* dataset.

guesses for each case), for each available view. This enables the system to run in a very short time, even though the processing pipeline is not trivial, and the data structures used have a high degree of complexity.

The processing time needed for each frame depends on the number of people detected: in order to measure the computational efficiency regardless of the complexity of the scene, the time needed for evaluating the pose of a person seen in two views was evaluated. Experiments were run on a Intel Core i7-4702HQ CPU running at 2.20 GHz; the measured average running time is 3.95 ms, ranging between 1.4 ms and 8.1 ms. The processing time is rather variable because it depends on the number of body poses with high confidence provided by the skeletal tracker, which in turn determines the number of points to be triangulated. The running time is nevertheless extremely low in any case considered, which demonstrates the high efficiency of the proposed method.

The values reported above do not include the time needed by the body pose estimation algorithm working on the single images, which is in average 0.82 s for the PBD algorithm working on each image of the campus dataset. The PBD algorithm did not saturate all the computational power of the CPU, which is made of four hyperthreading cores, supporting up to 8 threads; therefore, we

also measured the time needed for processing the three frames acquired from the three different viewpoints in parallel. In this second case, the CPU is saturated, and the average time needed for processing all three frames is 1.82 s. These measurements were obtained considering frames 200-400 of the dataset, which represent an average working condition (one or two people are found in the scene). The processing time of the PBD do not strongly depend on the framed scene, but it strongly depends on the image size and on the skeletal tracking algorithm employed, which can be changed thanks to the modularity of the proposed system. The time needed for evaluating the SPS is as reported in [5].

5. Conclusions

A system for people re-identification based on a camera network was presented. The re-identification is performed on the data provided by a skeletal tracker, an approach that already demonstrated very high performance on RGB-D sensory data. In this paper, the application of this idea to a new scenario was performed: switching from one single RGB-D sensor to a camera network imposed a completely different approach to body pose estimation, which can still be considered as an open problem in computer vision.

A novel, geometry-based approach was introduced for merging together the body pose detections obtained from the different viewpoints in an efficient way; the system is also able to match the detections when multiple targets are seen at the same time. The proposed approach was tested on a publicly available dataset, demonstrating that a sensible performance increase can be obtained thanks to the multi-viewpoint approach.

The system presented here is a first step towards a common framework capable of letting re-identification processes based on RGB-D and multi-viewpoint RGB sensors communicate. This can be achieved by designing similar target models that can be generated starting from any of the mentioned sensory systems. Further work will focus on such common target models, and on further improvements of the SPS.

Acknowledgements

This work was partially funded by the "Assegni di ricerca senior" funding scheme of the University of Padova, for the project "A unified approach for three-dimensional target modeling and tracking for camera networks and range sensors".

References

[1] R. B. Rusu, S. Cousins, 3D is here: Point Cloud Library (PCL), in: International Conference on Robotics and Automation (ICRA) 2011, Shanghai, China, 2011, pp. 1–4.

[2] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (12) (2013) 2878–2890.

[3] S. Zennaro, M. Munaro, S. Milani, P. Zanuttigh, A. Bernardi, S. Ghidoni, E. Menegatti, Performance evaluation of the 1st and 2nd generation kinect for multimedia applications, in: Multimedia and Expo (ICME), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–6.

[4] J. Shackleton, B. VanVoorst, J. Hesch, Tracking people with a 360-degree lidar, in: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, IEEE, 2010, pp. 420–426.

[5] M. Munaro, S. Ghidoni, D. T. Dizmen, E. Menegatti, A feature-based approach to people re-identification using skeleton keypoints, in: Robotics and Automation (ICRA), 2014 IEEE International Conference on, IEEE, 2014, pp. 5644–5651.

[6] S. Amin, M. Andriluka, M. Rohrbach, B. Schiele, Multi-view pictorial structures for 3d human pose estimation, in: British Machine Vision Conference, Vol. 2, BMVA Press, 2013.

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

[8] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification., in: British Machine Vision Conference, Vol. 2, 2011.

[9] K. Jungling, M. 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), 2009, pp. 30–37.

[10] K. Yoon, D. Harwood, L. Davis, Appearance-based person recognition using color/path-length profile, Journal of Visual Communication and Image Representation (JVCIR 2006) 17 (3) (2006) 605–622.

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

[12] L. Hu, S. Jiang, Q. Huang, W. Gao, People re-detection using adaboost with sift and color correlogram, in: Proc. of the 15th International Conference on Image Processing (ICIP 2008), 2008, pp. 1348–1351.

[13] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297–1304.

[14] S. Asteriadis, A. Chatzitofis, D. Zarpalas, D. S. Alexiadis, P. Daras, Estimating human motion from multiple kinect sensors, in: Proceedings of

the 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, ACM, 2013, pp. 1–6.

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

[16] J. Oliver, A. Albiol, A. Albiol, 3d descriptor for people re-identification, in: Proceedings of the 21st IEEE International Conference on Pattern Recognition (ICPR 2012), 2012, pp. 1395–1398.

[17] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, V. Murino, Re-identification with rgb-d sensors, in: European Conference on Computer Vision (ECCV) Workshops 2012, Springer, 2012, pp. 433–442.

[18] M. Munaro, A. Basso, A. Fossati, L. Van Gool, E. Menegatti, 3D reconstruction of freely moving persons for re-identification with a depth sensor, in: IEEE International Conference on Robotics and Automation (ICRA), Hong Kong (China), 2014, pp. 4512–4519.

[19] M. Munaro, A. Fossati, A. Basso, E. Menegatti, L. Van Gool, Person Re-Identification, Springer London, London, 2014, Ch. One-Shot Person Re-identification with a Consumer Depth Camera, pp. 161–181.

[20] M. Dantone, J. Gall, C. Leistner, L. Van Gool, Body parts dependent joint regressors for human pose estimation in still images, Pattern Analysis and Machine Intelligence, IEEE Transactions on 36 (11) (2014) 2131–2143.

[21] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, 2014, pp. 3686–3693.

[22] R. Hartley, A. Zisserman, Multiple view geometry in computer vision, Cambridge university press, 2003.

[23] N. Bellotto, H. Hu, Computationally efficient solutions for tracking people with a mobile robot: an experimental evaluation of bayesian filters, Autonomous Robots 28 (2010) 425–438.

[24] P. Konstantinova, A. Udvarev, T. Semerdjiev, A study of a target tracking algorithm using global nearest neighbor approach, in: CompSysTec 2003: e-Learning, ACM, 2003, pp. 290–295.

[25] J. Berclaz, F. Fleuret, E. Türetken, P. Fua, Multiple object tracking using k-shortest paths optimization, Pattern Analysis and Machine Intelligence, IEEE Transactions on 33 (9) (2011) 1806–1819.

[26] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic, 3D pictorial structures for multiple human pose estimation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1669–1676.

[27] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter, C. Pantofaru, An adaptable system for rgb-d based human body detection and pose estimation, J. Vis. Comun. Image Represent. 25 (1) (2014) 39–52.

[28] M. Munaro, E. Menegatti, Fast rgb-d people tracking for service robots, Autonomous Robots 37 (3) (2014) 227–242.

[29] M. Munaro, F. Basso, E. Menegatti, Openptrack: Open source multi-camera calibration and people tracking for rgb-d camera networks, Robotics and Autonomous Systems 75, Part B (2016) 525–538.