Multi-modal Person Re-Identification Using RGB-D Cameras

Federico Pala, Member, IEEE, Riccardo Satta, Giorgio Fumera, Member, IEEE, and Fabio Roli, Fellow, IEEE

Abstract—Person re-identification consists of recognizing individuals across different sensors of a camera network. Whereas clothing appearance cues are widely used, other modalities could be exploited as additional information sources, like anthropometric measures and gait. In this work we investigate whether the re-identification accuracy of clothing appearance descriptors can be improved by fusing them with anthropometric measures extracted from depth data, using RGB-D sensors, in unconstrained settings. We also propose a dissimilarity-based framework for building and fusing multi-modal descriptors of pedestrian images for re-identification tasks, as an alternative to the widely used score-level fusion. The experimental evaluation is carried out on two data sets including RGB-D data, one of which is a novel, publicly available data set that we acquired using Kinect sensors. The fusion with anthropometric measures increases the first-rank recognition rate of clothing appearance descriptors up to 20%, whereas our fusion approach reduces the processing cost of the matching phase.

Index Terms—Multi-modal person re-identification, Clothing appearance, Anthropometric measures, RGB-D sensors.

I. INTRODUCTION

PERSON re-identification consists of matching individuals across different, possibly non-overlapping views of a camera network [1]. It can enable various applications, like off-line retrieval of video sequences containing an individual of interest whose image is given as a query, and on-line pedestrian tracking (aka re-acquisition [2]). Strong biometric traits, like faces, cannot be exploited in typical settings characterized by strong pose variations, partial occlusions, low resolution and unconstrained environment [1] (see Fig. 1). Clothing appearance is the most widely used “soft”, session-based cue, since it is relatively easy to extract, and exhibits uniqueness over limited periods of time. The re-identification accuracy it can attain is however affected by low inter-class variability (i.e., different individuals wearing similar clothing), especially in scenarios involving a large number of people. Re-identification based only on clothing appearance can actually be difficult also for human operators, as pointed out, e.g., in [3]. For this reason some authors proposed to combine it with other modalities, like anthropometric measures [4], [5], gait [6] and thermal data [5], and reported some evidences that the proposed multi-modal systems can outperform systems based on clothing appearance alone.

In this work we focus on anthropometric measures, since their extraction has been considerably eased by recently introduced RGB-D sensors also in unconstrained re-identification settings, whereas RGB sensors require complex calibration procedures, and are very sensible to occlusions, clutter and lighting conditions. We address in particular two issues, that have not been considered in depth in previous work. (1) Can the re-identification accuracy of clothing appearance be improved by fusing it with anthropometric cues, in unconstrained re-identification settings? We address this issue by selecting anthropometric measures that can be extracted in such settings, using commercial RGB-D sensors like the Kinect, among the ones proposed in previous work; we then fuse them with different state-of-the-art clothing appearance descriptors. (2) How to combine descriptors coming from different modalities? All previous works used score-level fusion rules. Here we also explore a different possibility, by developing a fusion method based on feature-level fusion, extending our previously proposed Multiple Component Dissimilarity (MCD) descriptor [7]. We originally developed MCD for reducing matching time of clothing appearance descriptors; here we show that it exhibits some interesting properties for multi-modal fusion. As a by-product, we acquired a novel, publicly available data set of video sequences of individuals, exhibiting different poses and illumination conditions, using Kinect RGB-D sensors.

The paper is structured as follows. In Sect. II we survey existing re-identification techniques based on clothing appearance, and on multiple modalities, focusing on anthropometric measures. In Sect. III we describe our dissimilarity-based framework for multi-modal descriptors. In Sect. IV we choose a set of anthropometric measures that can be extracted from RGB-D data in unconstrained settings, and describe the clothing appearance descriptors used in the experiments. Experimental results are reported in Sect. V. Conclusions and suggestions for future research directions are given in Sect. VI.

II. PREVIOUS WORK

In this section we overview person re-identification methods based on clothing appearance (which is the only cue used in most of the existing methods), and on their combination with other modalities, focusing on anthropometric measures.
A. Person re-identification based on clothing appearance

Most of the existing descriptors are based on a multiple part - multiple component (MPMC) representation: they subdivide the body into several parts to deal with its non-rigid nature, and represent each part as a set of components using various kinds of local or global features. SDALF [8] is a paradigmatic example: it subdivides the body into left and right torso and legs, according to its symmetry and anti-symmetry properties. Three kinds of features are extracted from each part: color histograms in the HSV color space; maximally stable color regions (MSCR); recurrent high-structured patches (RHSP) (see Sect. IV-B). To extract MSCR and RHSP, several image patches are randomly sampled, and then clustered to find the most significant ones. In [9] the body is subdivided into head, torso and legs as in [8]. Each part is described using weighted Gaussian color histograms features (to capture the chromatic content of the region around SIFT key-points), pyramid of histograms of orientation gradients (by concatenating the histogram of gradients along edge lines), and Haralick features (to describe textures, based on the gray level co-occurrence matrix). In [10] the body is subdivided into upper and lower parts: each part is represented using the MPEG7 dominant color descriptor, and a learning algorithm is used to find the most discriminative appearance model. Our MCMimpl [11] subdivides the body into torso and legs, randomly extracts from each part rectangular, possibly overlapping patches, and represents them with HSV color histograms; to attain robustness to illumination changes, synthetic histograms are generated from the original ones by varying brightness and contrast. In [12] dense color histograms in several colour spaces, and different texture features, are extracted from four body parts (upper and lower torso, upper and lower legs); a nonlinear warp function between features from two cameras is learnt, to deal with large changes in appearance between them, due to different lighting conditions and poses, occlusion, and background clutter. In [13] the body is subdivided into six horizontal strips, and clothing appearance is modelled in terms of color and texture features as a function of the pose; a classifier is trained offline to identify the most discriminative features, and subject-discriminative features are further learnt online. In [14] a spatial pyramid is built by dividing an image into overlapping horizontal stripes of 16 pixels height; color histograms and histograms of oriented gradients are computed for each strip. A ranking method specific to re-identification, based on sparse discriminative classifiers, is also proposed in [14].

Other methods exploit more refined subdivisions. In [15], a body part detector is used to find fifteen non-overlapping square cells, corresponding to “stable regions” of the silhouette, which are represented by a covariance descriptor in terms of color gradients. Color histogram equalization is performed to improve robustness to changing lighting conditions. Descriptor generation and matching are performed through a pyramid matching kernel. The subdivision of [16] is based on decomposable triangulated graphs, and each part is described by color and shape features. Pictorial structures are used in [3] to detect chest, head, thighs and legs, which are described by HSV histograms and MSCR patches as in [8].

Other approaches treat the body as a whole, and represent it using various kinds of features: Haar-like features [10]; SIFT-like interest points [16], [17], [18]; texture (Schmid and Gabor filters) and color (histograms in different color spaces) [19]; global color descriptors (histograms, spatigrams, color/path-length) [20]; 4D multi color-height histograms and transform-normalized RGB (illumination-invariant) [21]; biologically-inspired features and covariance descriptors capturing shape, location and color information [22] (see Sect. IV-B).

The use of RGB-D sensors has recently been proposed, since they enable a more effective foreground/background segmentation and body part localization, with respect to RGB sensors [23], [24]. In [23] a “body print” is extracted from each individual, exploiting the precision of RGB-D sensors in measuring world coordinates corresponding to the pixels of the scene; the body is subdivided into horizontal stripes at different heights, and each stripe is represented as its mean RGB value along a video sequence. In [24] the Kinect SDK is used for extracting the torso and legs body parts, which are then represented using the same descriptor of [7].

B. Person re-identification using anthropometric measures

Anthropometry is the characterization of individuals through the measurement of physical body features, e.g., height, arm length, and eye-to-eye distance [25], typically taken with respect to landmark points like elbows, hands, knees and feet, which are localized automatically or manually. Their discriminant capability was discussed in a classic study [26], where ten different measures were evaluated over 4,063 individuals.

Although the use of anthropometric measures has already been proposed for personal identity recognition, existing methods do not fit the typical setting of person re-identification, i.e., multiple, usually non-overlapping cameras and unconstrained environments, with large pose variations and non-cooperative users. In some works, anthropometric measures are extracted using costly 3D devices, like 3D laser scanners, and require the user collaboration in a constrained setting [27], [28], [29]. Other methods use an RGB camera with no specific calibration [30], [31]; however, anthropometric measures can be evaluated in [30] up to a scale factor only, and thus cannot be used for comparing individuals in images acquired by different cameras, whereas in [31] images in frontal pose are required, and thirteen body landmarks have to be manually selected. Other methods estimate absolute height values only, but require ad hoc camera calibration [32], [33], [34], [35].
The use of anthropometric measures for re-identification was first proposed in [36], where height was estimated from RGB cameras as a cue for associating tracks of individuals coming from non-overlapping views (this corresponds to the re-acquisition task that is enabled by person re-identification), but ad hoc camera calibration is required. The extraction of anthropometric measures in re-identification settings has recently been made viable by novel RGB-D sensors, which provide a reliable, real-time body pose estimation [37], [38]. For instance, the Kinect SDK\textsuperscript{1} provides the absolute position of twenty body parts (see Fig. 2). This was exploited in [39] to extract different anthropometric measures from front and rear poses: distance between floor and head, ratio between torso and legs, height, distance between floor and neck, distance between neck and left shoulder, distance between neck and right shoulder, and distance between torso centre and right shoulder; three other geodesic distances are estimated from the 3D mesh of the abdomen, using the Kinect depth map: torso centre to left shoulder, torso centre (located in the abdomen) to left hip, and distance of torso centre to right hip. In [40] a specific setting was considered, in which the cameras are installed on the floor after an entrance door. The proposed anthropometric measures (extracted from a sequence of frames) are the individual’s average blob height, area and projected volume to the floor, and blob speed. Two kinds of descriptors have recently been considered in [41], exploiting Kinect sensors: 13 anthropometric measures extracted from the body joints, and a point cloud model of human body. We finally mention that anthropometric measures extracted from RGB-D sensors have been recently proposed also for person recognition tasks; e.g., height, arm and leg length were used in [42] and [43] for gait recognition.

We point out that some of the measures of [39] are difficult or even impossible to extract from unconstrained poses: measures from 3D mesh require near-frontal pose (the abdomen is hidden in rear pose); neck distance to left and right shoulders is hard to compute from lateral pose, even using a depth map, and requires to distinguish between left and right body parts. The measures used in [40] are tailored to the specific setting of top-camera views, instead, and cannot be used in standard re-identification settings. The skeleton measures of [41] are in principle suited to standard settings, and can be extracted from unconstrained poses.

C. Multi-modal person re-identification

Some authors recently proposed to combine clothing appearance with cues coming from other modalities: anthropometric measures, thermal features and gait. In [5] clothing appearance (RGB histograms from upper and lower body) was combined with thermal features (SURF descriptors), and anthropometric measures (averaged over different frames from the depth map): frontal curve model, encoding the distances from head to neck and neck to torso along the body surface; thoracic geodesic distances; and the length of seven inter-joint segments connecting different body parts. In [4] clothing appearance (color histogram of upper body and legs) was combined with the subject’s height, computed by subtracting the $y$-values in world-coordinates of the upmost and the lowermost silhouette points. In [6] clothing appearance (color histograms) was combined with gait, described by a spatio-temporal histogram of oriented gradients. In [44] clothing appearance (color histogram of head, torso and legs, and a texture model based on Local Binary Patterns) was combined as in [4] with the height, which was however extracted using only RGB sensors, by converting the detected part part positions in real world coordinates through a specific camera calibration.

We point out that in these works the improvement in re-identification accuracy that can be attained by the additional modalities over clothing appearance has not been clearly evaluated. Moreover, in all the existing methods where multi-modal cues are used, the different modalities are combined through score-level fusion techniques. A different fusion technique between different descriptors was proposed in [45], although it was evaluated only on clothing appearance cues: it exploits a multi-view, semi-supervised learning framework with manifold regularization in vector-valued Reproducing Kernel Hilbert Spaces; the similarity between two individuals is computed for each feature (descriptor) using a kernel function, and a learning algorithm is used to combine all the features, defining a mapping between their values and the identity of the individual of an input image, for a given template gallery.

III. DISSIMILARITY-BASED DESCRIPTORS FOR MULTI-MODAL PERSON RE-IDENTIFICATION

Here we summarize our MCD descriptor [7], and show how it can be extended to multi-modal re-identification.

A. Multiple Component Dissimilarity representation

Dissimilarity representations have been originally proposed as an alternative to feature vector representations, for pattern recognition problems where features are not available or it is difficult to find discriminative ones, whereas a dissimilarity measure between object pairs can be defined [46]. Objects are represented as vectors of dissimilarity values from a predefined set of “prototypes”, which can be chosen depending on the task.
at hand, e.g., by clustering [46], [47]. Dissimilarity representations are also used in computer vision tasks together with part-based object models, and/or local features (e.g., SIFT points), by defining prototypes representative of local object parts [48]. In [7] we exploited them to devise a framework for MPMC clothing appearance descriptors, in which (see Sect. II-A): (i) A body part subdivision is often used, and a distinct representation is built for each part (otherwise, the whole body can be viewed as a single component); (ii) Each part (or the full body, if no part subdivision is used) is represented by multiple components, e.g., patches, strips, or interest points, and each component is described with a distinct feature vector. The MPMC representation can be formalized as follows. The image $I$ of an individual is represented as an ordered sequence of descriptors $\{I_1, \ldots, I_M\}$ of $M \geq 1$ predefined body parts. Each descriptor is a set of feature vectors extracted from $n_m$ different components, $I_m = \{x_{m,1}, \ldots, x_{m,n_m}\}$, $x_{m,k} \in \mathcal{X}_m$ (different feature spaces $\mathcal{X}_m$ can also be used for different parts). An example is shown in Fig. 3.

Our MCD framework was originally aimed at speeding up matching of MPMC descriptor pairs, by converting them into dissimilarity vectors. This is obtained in two steps. **Prototype selection:** a “visual” prototype $P_m$ is defined off-line for each body part, from a given set $\mathcal{I}$ of images of individuals. Each prototype is obtained by merging the components of the $m$-th body part of all images in $\mathcal{I}$ and grouping them into $N_m$ clusters: $P_m = \{P_{m,j}\}_{j=1}^{N_m}$. Each cluster is thus a set of components, $P_{m,j} = \{p_{m,j}\}_{i=1}^{N_{m,j}}$, $p_{m,j} \in \mathcal{X}_m$, and represents a specific low-level visual characteristic of the corresponding part in the feature space $\mathcal{X}_m$ (e.g., a certain color distribution). **Dissimilarity descriptor computation:** from the original MPMC descriptor of any individual’s image $I$, a dissimilarity vector $D_m^I$ is obtained for each part $m$ as:

$$D_m^I = [d(I_m, P_{m,1}), \ldots, d(I_m, P_{m,N_m})]$$

where the superscript $D$ denotes a dissimilarity descriptor, and $d(\cdot, \cdot)$ is a dissimilarity measure between sets of components (e.g., the Hausdorff distance [49]); the vectors $D_m^I$ are then concatenated into a dissimilarity vector for the whole image:

$$D^I = [D_1^I, \ldots, D_M^I].$$

Computing the similarity between the MCD descriptors of of a probe and a template image amounts to comparing two real-valued vectors, which can be much faster than evaluating the similarity between the original descriptors [7]. In [7] we devised a similarity measure suitable to MCD descriptors, with the following rationale: if the images of two individuals $I'$ and $I''$ do not exhibit the local characteristic associated with a set of components $P_{m,j}$ (i.e., both $d(I'_m, P_{m,j})$ and $d(I''_m, P_{m,j})$ exhibit high values), then $P_{m,j}$ does not provide any information about their similarity. Conversely, the smaller the values of either $d(I'_m, P_{m,j})$ or $d(I''_m, P_{m,j})$, or both, the higher the information $P_{m,j}$ conveys about the similarity between $I'$ and $I''$. Accordingly, we defined a weighted Euclidean distance that gives higher weights to elements that exhibit smaller values either in $D$ or in $D''$. Assuming that $d(\cdot, \cdot) \in [0, 1]$:

$$D(D, D'') = \frac{\sum_{m=1}^{M_n} \sum_{j=1}^{N_m} w_{m,j} (d(I'_m, P_{m,j}) - d(I''_m, P_{m,j}))^2}{W},$$

where $w_{m,j} = (1 - \min\{d(I'_m, P_{m,j}), d(I''_m, P_{m,j})\})^2$, and $W$ is a normalization factor such that $\frac{1}{W} \sum_{m=1}^{M_n} \sum_{j=1}^{N_m} w_{m,j} = 1$ (in [7] a different definition of $w$ was used; we found that the one of Eq. (3) is more effective for multi-modal descriptors).

Finally, in [7] we showed that prototypes can be constructed off-line, even from a different set of images than the template gallery, without affecting re-identification accuracy. Furthermore, since prototype construction is an unsupervised procedure (i.e., the identity of the individual is not used), in off-line re-identification settings prototypes can be constructed using also the available probe images.

B. MCD descriptors for multi-modal re-identification

Here we show that our MCD framework can be applied to descriptors of other modalities beside clothing appearance, and that it provides an interesting solution for fusing descriptors of different modalities in a feature-level fashion. To this end, a given descriptor has first to be framed as a MPMC one. In particular, a given set of $q$ anthropometric measures can be seen as the simplest MPMC representation made up of one part (the whole body, $M = 1$) and one component, the feature vector $x_{1,1} \in \mathcal{X}_1 \subset \mathbb{R}^q$, so that the corresponding descriptor of an individual $I$ is given by $I_1 = \{x_{1,1}\}$. According to Sect. III-A, to obtain an MCD descriptor one has to construct one set of prototypes $P_1$, which is made up of a set of $N_1$ clusters, $P_1 = \{P_{1,j}\}_{j=1}^{N_1}$, obtained by grouping the vectors of anthropometric measures of a given set of individuals $\mathcal{I}$; each cluster is a set of components $P_{1,j} = \{x_{1,j}\}_{j=1}^{N_{1,j}}$, where each $x_{1,j} \in \mathcal{X}_1$ is the vector of anthropometric measures of a given individual in $\mathcal{I}$. The corresponding MCD descriptor will be given by $D^I = [D_1^I, \ldots, D_M^I]$. In general, given $K > 1$ different modalities and their original descriptors, one obtains $K$ dissimilarity vectors, each defined as in Eq. (2).

Consider now the issue of how to combine the descriptors of $K$ different modalities. In re-identification tasks one can use either feature-level fusion (concatenating the feature vectors of each modality into a single one, and then computing an overall matching score), or score-level fusion (fusing the matching scores computed separately for each modality). As pointed out in Sect. II-C, in all previous works on multi-modal re-identification score-level fusion was used, although the fusion
method of [45] can also be used for this purpose. Score-level fusion appears as the most straightforward solution, for at least two reasons. One reason is that in multi-modal systems feature-level fusion requires one to concatenate heterogeneous quantities, like a color histogram for clothing appearance and a set of anthropometric measures (the same issue arises, e.g., in multi-modal biometric identity recognition [50]), whereas score-level fusion allows one to combine homogeneous information, i.e., a set of similarity scores (one for each modality). Another reason is that descriptors that lie in high-dimensional feature spaces (like many clothing appearance ones) may overwhelm the contribution of descriptors that lie in relatively lower dimensional spaces (e.g., a vector made up of a few anthropometric measures). Our dissimilarity-based MCD descriptor provides however a different perspective for feature-level fusion. First, dissimilarity-based descriptors are representation-independent, i.e., they are logically and semantically at a higher level than the underlying object representation. This implies that dissimilarity values computed on different modalities are semantically as coherent as the matching scores computed from the original descriptors: in the case of MCD, the only difference is that they encode a (dis)similarity between local object components (e.g., body parts) instead of between whole objects. Second, the size of MCD descriptors can be controlled by setting the desired number of prototypes (see Sect. III-A), independently of the feature set size of the original descriptors; this allows one to avoid concatenating vectors of very different size. Obviously, reducing the number of prototypes below a certain amount may affect the resulting re-identification accuracy: this issue will be empirically investigated in Sect. V. To sum up, feature-level fusion of MCD descriptors is in principle not affected by the issues that affect non-dissimilarity descriptors.

Our MCD framework can be extended to \( K \) different modalities as follows. A distinct prototype set is first constructed for each of modality. Then, given the image of an individual and the \( K \) original descriptors, the corresponding dissimilarity vectors \( \mathbf{I}_{m}^{D,k}, k = 1, \ldots , K \), are computed, and the final MCD descriptor is obtained by concatenating them in any predefined order, e.g.: \[
\mathbf{I}_{m}^{D} = [\mathbf{I}_{m}^{D,1}, \ldots , \mathbf{I}_{m}^{D,K}].
\] (4)

The matching score between two such descriptors can finally be computed using again Eq. (3). The proposed multi-modal MCD representation is summarized in Fig. 4.

IV. CLOTHING APPEARANCE DESCRIPTORS AND ANTHROPOMETRIC MEASURES

To investigate the two issues mentioned in Sect. I, here we explain our choice of a set of anthropometric measures that can be extracted from RGB-D sensors from unconstrained poses, and of clothing appearance descriptors. We also describe the construction of the corresponding MCD descriptors.

A. Anthropometric measures

The depth map and the person detection functionality provided by off-the-shelf RGB-D sensors enable a relatively easy detection of some anthropometric measures (see Sect. II-B). Further ones can be extracted exploiting the pose estimation functionality of the Kinect SDK (based on converting depth data to cloud points in real-world coordinates), which provides a real-time estimation of the absolute position in metric coordinates of 20 different body joints (see Fig. 2): the spine, the centre of the hip, shoulder and head, left and right shoulder, elbow, wrist, hand, hip, knee, ankle and foot. This allows further measures to be extracted. Each joint is associated with a tracking state: “tracked”, if it is clearly visible, and “inferred” if it is not, but the Kinect SDK can infer its position. The joints that can be actually tracked or inferred, and the precision of their localization, depend on the pose; e.g., in a lateral pose the joint of the farthest shoulder is usually not tracked.

To the purpose of this work, we selected a set of anthropometric measures among the ones proposed in previous work (see Sect. II-B), focusing on measures that can be extracted from unconstrained poses, and with low processing cost, to fit real-world video surveillance and re-identification scenarios. For instance, this is the case of the height of a person: it can be extracted from the silhouette, e.g., by measuring the distance between the highest silhouette point and the floor plane, in real-world coordinates. This is not the case of geodesic distances of [39]: they were estimated from the 3D mesh of the abdomen, which can be extracted only from a frontal pose, and with a relatively higher complexity. Specific issues also arise for measures that can be extracted from skeleton joint positions. The positions of some joints are estimated more reliably from frontal poses (probably because the Kinect device has been designed for tracking individuals standing in front of the sensor), and may change significantly across different poses; e.g., in a rear pose the head and the neck joint are usually localized in a higher position than in a frontal pose (see Fig. 2(d)). Moreover, although the distances between all pairs of adjacent joints could be used as anthropometric measures, some pairs of joints are closer than others (e.g., the hip width and the shoulder width), and thus their estimated distance may be affected by a higher relative error and exhibit a higher variance. Accordingly, we selected the following seven measures from [39] and [41] (denoted as \( d_{1} \) to \( d_{7} \)), and two measures (\( d_{8} \) and \( d_{9} \)) from [42]:

- \( d_{1} \) distance between floor and head
- \( d_{2} \) ratio between torso and legs
- \( d_{3} \) height (distance between the highest body silhouette point and the floor plane)
- \( d_{4} \) distance between floor and neck
- \( d_{5} \) distance between neck and shoulder
- \( d_{6} \) distance between torso centre and shoulder
- \( d_{7} \) distance between torso centre and hip
- \( d_{8} \) arm length (sum of the distances between shoulder and elbow, and between the elbow and wrist)
- \( d_{9} \) leg length (sum of the distances between hip and knee, and between knee and ankle)

All distances are Euclidean. In particular, \( d_{6} \) and \( d_{7} \) were computed as geodesic distances in [39], but we replaced them with Euclidean ones to make them pose-invariant; we also averaged the pairs of measures exhibiting vertical symmetry.
(d₅, d₆, and d₇), if both are in the “tracked” status; otherwise we used only the “tracked” one. Note also that we computed d₂ as in [39]:
\[ d₂ = \frac{d₅}{d₁ \cdot d₆ \cdot d₇} \]
We finally normalized all these measures (both in template and in probe images) to zero mean and unit variance (mean and variance were computed on the template gallery).

According to Sect. III-B, we built a MPMC descriptor of anthropometric measures made up of one body part and one component. The latter is represented as a vector \( x = [d₁, \ldots, d₅] \). In [39] each value was computed from the video frame exhibiting the highest number of joints with status “tracked”. To improve robustness, we computed each value as the median over the first ten frames from a video. We will compare the two strategies in Sect. V-B. We then computed the matching score \( s \) between the descriptors of two individuals \( x' \) and \( x'' \) as in [39], using a weighted Euclidean distance to take into account the different discriminant capability of each measure:
\[ s = \sum \limits_k w_k (d'ₖ - d''ₖ)^2, \tag{5} \]
with \( w_k \geq 0 \) and \( \sum w_k = 1 \). Details about weight computation are given in Sect. V-B.

### B. Clothing appearance descriptors

We chose three MPMC clothing appearance descriptors: the state-of-the-art SDALF [8] and eBiCov [22], and our MCMimpl [11], which we used in our first work on the MCD framework [51]. We implemented SDALF and eBiCov using the source code provided by the authors. SDALF subdivides the body into torso and legs through a horizontal axis that is found by exploiting symmetry and anti-symmetry properties of the silhouette’s color and shape, and used three kinds of features. Maximally Stable Color Regions (MSCR) are non-regular regions of homogeneous color, extracted from the whole body, which describe the per-region color displacement, and are found via agglomerative clustering; each one is represented by its area, centroid, second moment matrix and average color, resulting in a 9-dimensional vector. Recurrent High-Structured Patches (RHSP) are rectangular patches made up of recurrent, repeated patterns, separately extracted from each part, and represented by a rotation-invariant LBP histogram; they highlight texture characteristics that are highly recurrent in the pedestrian appearance. Both MSCR and RHSP are sampled mainly around the vertical axis of symmetry of each body part. Weighted HSV histograms (w-HSV) are extracted from each body part to capture the chromatic content (giving lower weights to pixels closer to the body periphery), and are concatenated into a single feature vector. SDALF can be conveniently seen as being made up of \( M = 4 \) sets of components: the MSCR feature vector, the RHSP feature vectors extracted from torso and legs, and the concatenated HSV color histogram.

The eBiCov (“enriched gBiCov”) descriptor [22] combines SDALF with the gBiCov descriptor. gBiCov is made up of a Biologically Inspired Features (BI) [52] and a Covariance (COV) descriptor. Two layers were selected from BI: Gabor filters and the MAX operator, for improving respectively the robustness to illumination changes, and to scale changes and image shifts. COV is used to compute the similarity of BI features taken at neighboring scales, capturing shape, location and color information. Each of them is extracted from the whole body (without background subtraction) separately from the three HSV channels, and the three resulting feature vectors are then concatenated. Therefore, gBiCov can be seen as made up of two components (BI and COV) extracted from a single part, and eBiCov as a descriptor made up of \( M = 6 \) components (4 for SDALF and 2 for gBiCov).

The original MCMimpl descriptor uses the same body subdivision as SDALF. In this work we used an enhanced version, exploiting the skeleton points extracted by the Kinect SDK: using only points that can be detected from any pose (see Fig. 2(b),(c),(d)), we subdivided the body into \( M = 4 \) parts: upper and lower torso, upper and lower legs. The torso region is localized as the portion of the image between the \( y \) coordinates of shoulder and hip centers. The mask pixels corresponding to the first half of the torso region are considered as the upper torso, and the other ones as the lower torso. The mask pixels between the coordinate of the hip centre and the average of the \( y \) coordinates of the knees (or the \( y \) coordinate of the visible knee, if only one is detected), define the upper legs region. The mask pixels between the average \( y \) coordinate of the knees and the bottom of the mask define the lower leg region. The set of components of each body part is obtained by randomly extracting image patches of different...
sizes, possibly overlapping; each patch is represented with an HSV color histogram.

We refer the reader to [8], [11], [22] for further details.

C. Computing MCD descriptors

To obtain a MCD descriptor we had to choose a clustering technique for prototype construction, and a distance measure \( d(\cdot, \cdot) \) between sets of components (see Eq. 1). For non-singleton sets of components (e.g., the HSV histogram in MCMimpl), we used the two-stage clustering approach of [7], aimed at reducing the processing cost. It consists of a first Mean-Shift clustering step [53] applied to each set of components, and a subsequent c-Means step applied to the first-stage centroids. For singleton set of components (e.g., the vector of anthropometric measures), only the c-Means step was carried out. The prototypes were defined as the resulting \( c \) centroids. In our experiments we set the bandwidth parameter of Mean-Shift to 0.3, \( c = 200 \) for clothing appearance descriptors, and \( c = 30 \) for anthropometric measures. This choice of \( c \) is discussed in Sect. V, where we evaluate how the number of prototypes affects the re-identification accuracy.

We defined \( d(\cdot, \cdot) \) as the modified \( k \)-th Hausdorff distance [49], which is known to be robust to outliers. It is defined as the \( k \)-th ranked minimum distance between all pairs of elements from two sets \( P \) and \( Q 
\[
\begin{align*}
    d(P, Q) &= \max \{ h_k(P, Q), h_k(Q, P) \}, \\
    h_k(P, Q) &= k\text{-th} \min_{p \in P, q \in Q} (\|p - q\|).
\end{align*}
\]

The parameter \( k \) controls the influence of outliers. We set \( k = 10 \). We then chose the same distance metric \( \| \cdot \| \) between components, both for the SDALF and MCMimpl descriptors (we refer the reader to [8], [11] for further details). For the chosen anthropometric descriptor, the set of components is a singleton (a single feature vector). Hence, we defined \( \| \cdot \| \) as the weighted distance
\[
\| x' - x'' \| = \sum_k w_k (d'_k - d''_k)^2,
\]
using the same weights \( w_k \) of Eq. (5) to take into account the discriminant capability of the different measures.

V. Experimental evaluation

According to Sect. I, the goals of our experiments are the following: (1) Evaluating whether anthropometric cues can improve the re-identification accuracy of clothing appearance ones in unconstrained re-identification settings; to this end, we carried out experiments on the three clothing appearance descriptors mentioned above. (2) Evaluating two different techniques for combining multi-modal descriptors: score-level fusion, and our dissimilarity-based MCD feature-level fusion. We describe in Sect. V-A the data sets we used and the experimental setup, in Sect. V-B the weights assigned to each of the chosen anthropometric measure, and in Sect. V-C the experimental results.

A. Data set and experimental setup

To carry out our experiments, a data set including RGB and depth data (in particular, the estimated positions of the joints) is required. Most benchmark data sets for person re-identification were acquired using only RGB sensors, e.g., [54], [55], [56]. To our knowledge, the only data set that contains also RGB-D data and can be used for our purposes is “RGBD-ID” [39]; since it was designed mainly for using depth data (sometimes the same individual wears different clothes in different acquisitions), we modified it as described below. The BIWI RGBD-ID data set of [41] was designed for a long-term setting, and therefore most of the individuals wear different clothes in training and testing sequences; it is thus not suited to short-term re-identification settings including clothing appearance descriptors. Beside using the data set of [39], we also acquired a new data set of video sequences, named “KinectREID”, which is available upon request.2

Our KinectREID data set was acquired using Kinect sensors and the official Microsoft SDK. It consists of video sequences of 71 individuals taken at a lecture hall of our department, under different lighting conditions and three view points: three near-frontal views, three near-rear views, and one lateral view. All the individuals were requested to walk normally along a predefined path; some of them carried accessories like bags. Seven video sequences were taken for each individual, for a total of 483 video sequences (14 sequences showing lateral poses, on which the SDK tracking failed, were discarded). Each sequence lasts for about 10 sec., but depth data is available only for a few seconds, corresponding to the range of the Kinect device (about 0.8 to 4.0 m). Some examples are shown in Fig. 5. Each tracked individual was associated with a sequence of frames, with the corresponding segmentation masks and skeleton points. We used both the RGB frames and the skeleton points to extract one clothing appearance descriptor from each frame, and the skeleton points to estimate the anthropometric measures.

RGBD-ID [39] was acquired using Kinect cameras as well, but with the OpenNI SDK.3 It contains RGB and depth data for 80 individuals. Differently from KinectREID, in RGBD-ID: (i) the joint positions were obtained using another tracker, and wrist and ankle are not included; (ii) four acquisitions were made for each individual, one rear and three frontal poses, and in one of the latter the arms are stretched; (iii) for each acquisition only 4 or 5 RGB-D frames are provided; (iv) sometimes, the same individual wears different clothes in different acquisitions: we removed the corresponding tracks. Hence, 2 to 4 acquisitions remained for each individual, for a total of 197 video sequences out of the 320 original ones.

We carried out our experiments simulating a closed-set scenario, as in most of the existing works on person re-identification. In this scenario, the identity of each probe individual corresponds to one of the template identities. On both data sets, the experimental setup was the following. First, we selected 20 individuals for estimating the weights of the anthropometric measures (see Sect. V-B). These indi-

2 More information at http://prlab.diee.unica.it/en/PersonReIdentification
3 http://structure.io/openni
viduals were not used in the subsequent steps. For each of the remaining individuals, one video sequence was randomly chosen as a template track, and the remaining ones were used as probes. MCD prototypes were computed on the template gallery: $c = 200$ prototypes were used for clothing appearance descriptors, and $c = 30$ for anthropometric measures (see Sect. V-C). For matching a pair of probe and template tracks we used the multiple shots vs. multiple shots (MvsM) setup [8], i.e., we matched several pairs of the respective frames. To reduce processing cost we discarded from KinectREID the frames in which the skeleton was not available, and chose the first 10 remaining frames from each track. We used all the available frames (4 or 5) from RGBD-ID, instead. We evaluated the matching score between each pair of frames using Eq. (6) for the original descriptors, and Eq. (3) for MCD descriptors, and used their median value as the final matching score. The score-level fusion of the original clothing appearance descriptors was implemented as the widely used sum rule, which in preliminary experiments attained better results than other rules (minimum, maximum and product of the scores). We repeated the above procedure for ten times, and evaluated the results as the average Cumulative Matching Characteristics (CMC) curve: it is defined as the probability of finding the correct match within the first $n$ ranks, with $n$ ranging from 1 to the number of templates. All the frames used as probes and templates in each run, and the corresponding skeletal points, are available upon request (see the URL above).

B. Combination of the anthropometric measures

As explained in Sect. IV-A, to compute a matching score between two anthropometric descriptors (in our case, two vectors of anthropometric measures), and their similarity measure in the case of the corresponding MCD descriptor, we used a weighted combination of the normalized anthropometric measures, respectively Eq. (5) and (8). We computed the weights separately on the two data sets. To this end we maximized the AUC$_{20\%}$ performance index (the area of the first 20% ranks of the CMC curve, normalized to $[0,1]$) obtained by computing the matching score as in Eq. (5), on the subset of 20 individuals mentioned above. To find the “optimal” weights we used the quasi-exhaustive strategy of [39], i.e., a grid search in the weight space, considering for each weight the values from 0 to 1 with step 0.05. We also compared the two strategies mentioned in Sect. IV-A for computing the matching score: using only the pair of probe and template frames exhibiting the highest number of joints with status “tracked” as in [39], and computing the median over all the considered frames (10 on KinectREID, 4 or 5 on RGBD-ID).

In both data sets we obtained the highest AUC$_{20\%}$ value by computing the median score among all the available frames: 57% vs 43% on KinectREID, and 60% vs 57% on RGBD-ID (the lower improvement on RGBD-ID is probably due to the lower number of available frames). Accordingly, we used this strategy in the rest of the experiments. The weight values (see Sect. IV-A for the corresponding anthropometric measures) were the following. On KinectREID: $w_2 = 0.2$, $w_3 = 0.5$, $w_4 = 0.05$, $w_8 = 0.05$, $w_9 = 0.2$, whereas $w_1 = w_5 = w_6 = w_7 = 0$. Probably the distance between floor and head ($d_1$) was not discriminant, since the head joint in the rear pose is usually tracked in an higher position by the Kinect SDK, whereas the distances between near joints (neck to shoulders $d_5$, torso centre to shoulders $d_6$, and torso centre to hips $d_7$) are probably too small and therefore subject to higher relative errors. Some measures may also be redundant, due to a high correlation with other ones. On RGBD-ID we obtained the following weights: $w_1 = 0.4$, $w_3 = 0.6$, $w_8 = 0.05$ and zero for the remaining ones. The differences between the weights computed on the two data sets are due to the different acquisition settings, as well as on the different SDK (e.g., probably the distance $d_1$ between floor and head is computed differently in the Microsoft and OpenNI SDKs).

C. Experimental results

We first evaluate the discriminant capability of the anthropometric measures, both individually and jointly. For the $k$-th measure alone, the matching score was computed as $(d_k^p - d_k^t)^2$. For all measures, we computed the matching score using the weights reported in Sect. V-B, both for the original MPMC descriptor and for its MCD version. The CMC curves for both data sets are reported in Fig. 6. As one could expect, the individual anthropometric measures exhibit a very different range of performance. In particular, the height estimated as the distance between the highest body silhouette point and the floor plane ($d_3$) exhibited a very good performance on both data sets, if compared to the one of clothing appearance descriptors (see below, and Fig. 7). The combination of all anthropometric measures (the ones with non-zero weight) attained on both data sets a better performance than each individual one; sometimes the performance was similar or even better than the one of clothing appearance descriptors. Note also that the original descriptor and the MCD one exhibited a similar performance on RGBD-ID, and that the latter outperformed the former on KinectREID for ranks from about 5 to 20.

The CMC curves of the clothing appearance descriptors and of their fusion with the anthropometric ones are reported in Fig. 7 (the plots in each row correspond to one of the
clothing appearance descriptors, the plots in each column to one data set). We first compare the CMC of each of the three clothing appearance descriptors (either the original or the MCD one) with the corresponding multi-modal fusion with the anthropometric descriptor (i.e., each pair of lines of identical color in each plot of Fig. 7), which corresponds to the objective (1) mentioned at the beginning of this section. It can be seen that fusing the two modalities always produced a remarkable performance improvement over the clothing appearance descriptor alone, across a large range of ranks. The only exception is the MCMimpl descriptor on RGBD-ID, where the performance improvement is limited to the first few ranks, which are nevertheless the most relevant ones for re-identification tasks. This provides evidence that anthropometric measures actually provide complementary discriminant information with respect to clothing appearance cues. In particular, we observed that they allow discriminating between different template individuals wearing clothes similar to the probe: this is the reason of the improvement in recognition rate observed in all our experiments, already at rank 1.

We now compare the score-level fusion of the original descriptors with our feature-level MCD fusion technique (i.e., the pair of solid lines in each plot), which corresponds to our objective (2). Both fusion techniques attained a similar performance; the only exception is again MCMimpl on RGBD-ID, where the score-level fusion outperformed the MCD feature-level fusion for the lowest ranks. However, we point out that the MCD fusion technique has the advantage of a much lower processing cost for the matching phase of the clothing appearance component of the dissimilarity vector, as explained in Sect. III-A. This makes our MCD fusion technique suitable for real-time multi-modal person re-identification, e.g., on a laptop with a dual core i5-2410M processor, computing one descriptor from a single frame (including all preprocessing steps) took about 50 msee., whereas matching one probe with one template track took about 0.03 msee.

For the sake of completeness, we also compare the original clothing appearance descriptors and the corresponding MCD ones (i.e., the two dashed lines in each plot of Fig. 7). MCD descriptors of clothing appearance exhibited a similar (e.g., for SDALF and MCMimpl on KinectREID) or lower performance (SDALF and eBiCov on RGBD-ID) than the original ones. This is in agreement with our previous results [51]: in the case of clothing appearance cues, MCD descriptors can be useful to attain a trade-off between re-identification accuracy and processing cost; sometimes, they improve both.

We finally discuss how the size of prototypes affects the accuracy of MCD descriptors. The number of prototypes depends on the value of the \( c \) parameter of the \( c \)-Means clustering algorithm we used for prototype construction (see Sect. IV-C); obviously, their number also affects the processing time for dissimilarity computation [7]. To get a concise overview, we report in Fig. 8 the average AUC\(_{20\%}\) attained by MCD clothing appearance and anthropometric descriptors (solid lines) on KinectREID, as a function of \( c \). Similar results were observed on RGBD-ID. Note that the AUC\(_{20\%}\) of the anthropometric descriptor ends at \( c = 51 \): the reason is that no more than 51 prototypes can be obtained in MCD, since this MCD descriptor is made up of a single component and the number of template individuals is 51. For reference, the AUC\(_{20\%}\) values of the original descriptors (which do not depend on \( c \)) are also reported (dashed lines). For both modalities, the AUC\(_{20\%}\) initially grows as the number of prototypes increases, and attains a nearly constant value beyond a certain value of \( c \). This value is about 200 for all clothing appearance descriptors, and about 30 for the anthropometric descriptor. These are the values of \( c \) that we used in our experiments. These results suggest that a relatively small number of prototypes can provide a good trade-off between re-identification accuracy and processing time in our MCD descriptors.

VI. CONCLUSIONS

We investigated whether anthropometric measures can improve the re-identification performance of the widely used clothing appearance cue, in unconstrained settings, exploiting the depth information and the related functionality (in particular, the estimation of joint positions) provided by recently introduced RGB-D sensors. To this end we chose a subset of anthropometric measures proposed by other authors, which can be computed from unconstrained poses, and considered three different clothing appearance descriptors. The multi-modal fusion of the two cues always attained a better performance than the clothing appearance cue alone, providing evidence that anthropometric measures provide complementary discriminant
Fig. 7. CMC curves (left: KinectREID data set; right: RGBD-ID data set) attained by the three clothing appearance descriptors (from top to bottom: SDALF [8], eBiCov [22], MCMimpl [11]), in their original (dashed red lines) and MCD [7] version (dashed blue lines), and by their fusion with anthropometric descriptors (original descriptors: solid red lines; MCD descriptors: solid blue lines). MCD descriptors are denoted with the superscript “DIS”.

Fig. 8. Normalized AUC$_{20\%}$ as a function of prototype size $c$, attained on KinectREID by the MCD clothing appearance and anthropometric descriptors (solid lines). For reference, the AUC$_{20\%}$ of the original descriptors is also shown (dashed lines).

Information; in particular, they allow discriminating between template individuals wearing clothes similar to the probe. We also proposed a novel dissimilarity-based, feature-level fusion technique for multi-modal re-identification, based on our MCD descriptor previously proposed for clothing appearance, as an alternative to score-level fusion, which is the only technique used so far for multi-modal re-identification. We showed that our technique can attain a better trade-off between re-identification accuracy and processing cost, when complex descriptors are involved (like clothing appearance ones). As a by-product, we acquired a novel, publicly available data set of video sequences with Kinect sensors, including both RGB and depth data.

Several future research directions can be envisaged, in the context of multi-modal re-identification using RGB-D cameras, e.g.: (i) Investigating a wider range of anthropometric cues to further improve re-identification accuracy; (ii) Developing a framework that takes into account missing cues or modalities, due, e.g., to occlusions or to the pose of an individual; (iii) Experimentally comparing the fusion technique of [45] and our MCD-based technique; (iv) Investigating the use of other modalities beside clothing appearance and anthropometric measures, as well as the fusion of different descriptors of clothing appearance (which has been already addressed in [45]); e.g., skeleton-based gait [57] could be an effective cue, whose extraction is enabled as well by RGB-D sensors, whereas remote face recognition could provide some useful cues in the case when of both the template and the probe are in frontal pose [58].


Federico Pala received his M. Sc. degree in Electronic Engineering, with honours, from the University of Cagliari, Italy. Since 2013 he is a Ph.D. student in Electronic and Computer Engineering and since 2012, Research Fellow of the Pattern Recognition and Application laboratory (http://pralab.diee.unica.it/en/FedericoPala), both at the same University. His research interests are mainly related to computer vision, machine learning and pattern recognition techniques for smart video surveillance systems, in particular appearance-based person re-identification, and people retrieval based on textual attributes. He is a member of IEEE, of the IEEE SMC Society, and of the International Association for Pattern Recognition (IAPR), and is in the Executive Committee of the IEEE and SMC Student Branch of Cagliari, Italy.

Riccardo Satta is Scientific Officer at the Institute for the Protection and Security of the Citizen of the European Commission Joint Research Centre (JRC), and Research Fellow of the Pattern Recognition and Applications Lab of the University of Cagliari, Italy. He holds an MSc degree with full honours in Electronic Engineering, and a PhD in Information Engineering, both from the University of Cagliari, Italy. He conducts research mainly on the fields of forensic-targeted computer vision, digital image forensics, biometrics and machine learning.

Currently, he is principal researcher of the JRC project "Authors and Victims Identification of Child Abuse on-line", aimed at developing image and video analytics to help law enforcers in fighting against Child Abuse on-line. Before his current appointment, he was visiting researcher at the Digital Forensics Lab of the University of Warwick, UK, research fellow at the Ambient Intelligence Lab in Pula, Italy, and served as technical-field functionary for the local government of the Autonomous Region of Sardinia, Italy. He is member of the Europal Platform of Experts (EPE), of the International Association of Pattern Recognition (IAPR), and of the Institute for Systems and Technologies of Information, Control and Communication (INSTICC).

Giorgio Fumera received the M. Sc. degree in Electronic Eng., with honors, and the Ph.D. degree in Electronic Eng. and Computer Science, respectively in 1997 and 2002, from the University of Cagliari, Italy. Since February 2010 he is Associate Professor of Computer Eng. at the Dept of Electrical and Electronic Eng. of the same University. His research interests are related to methodologies and applications of statistical pattern recognition, and include multiple classifier systems, classification with the reject option, adversarial classification and document categorization. On these topics he published seventy papers in international journal and conferences. He acts as reviewer for the main international journals in this field, including IEEE Trans. on Pattern Analysis and Machine Intelligence, Journal of Machine Learning Research, and Pattern Recognition. He is a member of the International Association for Pattern Recognition, the IEEE Institute of Electrical and Electronics Engineers, and of the Italian Association for Artificial Intelligence (AI*IA).

Fabio Roli received his M.S. degree, with honours, and Ph.D. degree in Electronic Engineering from the University of Genoa, Italy. He was a member of the research group on Image Processing and Understanding of the University of Genoa, Italy, from 1988 to 1994. He was adjunct professor at the University of Trento, Italy, in 1993 and 1994. In 1995, he joined the Dept. of Electrical and Electronic Engineering of the University of Cagliari, Italy, where he is now professor of computer engineering and Director of the research lab on pattern recognition and applications (http://pralab.diee.unica.it). Dr Roli’s research activity is focused on the design of pattern recognition systems and their applications to biometric personal identification, multimedia text categorization, and computer security. He is a member of the governing boards of the International Association for Pattern Recognition and of the IEEE Systems, Man and Cybernetics Society. He is Fellow of the IEEE, and Fellow of the International Association for Pattern Recognition.