# A general method for appearance-based people search based on textual queries

Riccardo Satta, Giorgio Fumera, and Fabio Roli

Dept. of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{riccardo.satta,fumera,roli}@diee.unica.it

**Abstract.** Person re-identification consists of recognising a person appearing in different video sequences, using an image as a query. We propose a general approach to extend appearance-based re-identification systems, enabling also textual queries describing clothing appearance (e.g., "person wearing a white shirt and checked blue shorts"). This functionality can be useful, e.g., in forensic video analysis, when textual descriptions of individuals of interest given by witnesses are available, instead of images. Our approach is based on turning any given appearance descriptor into a dissimilarity-based one. This allows us to build detectors of the clothing characteristics of interest using supervised classifiers trained in a dissimilarity space, independently on the original descriptor. Our approach is evaluated using the descriptors of three different re-identification methods, on a benchmark data set.

## 1 Introduction

Person re-identification is a computer vision task for video-surveillance applications. It consists of recognising a person appearing in different video sequences taken by one or more cameras, using an image as a query. Since the face region has usually a small size, and people are often not in frontal pose, face recognition systems can not be applied. Thus, methods proposed so far exploit clothing appearance [1], or *soft* biometrics like gait [2]. In this paper we consider a similar task that we call "appearance-based people search". It consists of finding, among a set of images of individuals, the ones relevant to a *textual* query describing clothing appearance of an individual of interest. Thus, it differs from person re-identification, where the query is an *image* of the person of interest. This can be useful in applications like forensics video analysis, where a textual description of the individual of interest given by a witness can be available, instead of an image.

To our knowledge, an analogous task ("person attribute search") was considered so far only in [3, 4]. In [3] the basic idea of building a specific detector for each attribute of interest (e.g., the presence of beard and eyeglasses, the dominant colour of torso and legs, etc.), was proposed, and a specific implementation was developed, mainly for face attributes. The work in [4] focused on the following attributes: gender, hair/hat colour, clothing colour, and bag

(if any) position and colour, and a generative model was proposed to build the corresponding descriptors. Both works considered only torso and legs colour as clothing appearance attributes.

In this work, we propose instead a general approach to extend appearance-based person re-identification systems, exploiting the *same* descriptors of clothing appearance to enable also the people search functionality based on a textual query. Our approach relies on dissimilarity-based descriptors, which can be obtained using the Multiple Component Dissimilarity (MCD) framework of [5] from *any* appearance descriptor that uses a body part subdivision and a multiple instance representation. Such kind of descriptors is used in most of the current re-identification methods. In [5], MCD dissimilarity descriptors were exploited to speed up the task of person re-identification. In this paper we show that they can also be exploited to implement the appearance-based people search task. In this context, the advantage of dissimilarity descriptors is that they allow one to build detectors of the attributes of interest (e.g., the presence of a colour in the torso) using supervised classifiers, without requiring techniques tailored to the specific, original descriptors, as in the approaches of [3, 4].

The MCD framework is summarised in Sect. 2. Our approach to implement people search is described in Sect. 3, and is experimentally evaluated in Sect. 4 on a benchmark data set for person re-identification.

## 2 MCD-based appearance descriptors

The descriptors used by most appearance-based re-identification methods (1) subdivide human body into parts, and (2) represent each body part as a bag of low-level local features (e.g., random patches, or SIFT points) [6]. In [5] it was shown that any such descriptor can be turned into a dissimilarity one, which consists of a vector of dissimilarity values to a predefined set of visual prototypes. The aim of the MCD framework was to reduce processing time for real-time applications. In Sect. 3 we will show how MCD can also be exploited for the people search task. Here we summarise the procedure for building MCD descriptors.

A generic appearance descriptor $\mathbf{I}$ of an individual is a sequence $\{I_m\}_{m=1}^{M}$ of sets of "components", each one associated to one of the $M \geq 1$ body parts. Each $I_m$ is a bag of local feature vectors $\{\mathbf{i}_m^k\}_{k=1}^{n_m}$.

Let $\mathcal{I}$ be a *gallery* of appearance descriptors (see Fig. 1-a). To represent them in a *dissimilarity space* [7], a set of "visual" prototypes $\mathbf{P}_m = \{P_{m,p}\}_{p=1}^{N_m}$, is first constructed for each body part. Prototypes correspond to low-level visual characteristics (e.g., a certain distribution of colours) shared by several descriptors of $\mathcal{I}$. Then, for each $\mathbf{I} \in \mathcal{I}$, a dissimilarity descriptor $\mathbf{i}^{\mathrm{D}}$ is created, as a vector of dissimilarity values between each $I_m \in \mathbf{I}$, and the corresponding prototypes $\mathbf{P}_m$. Note that, contrary to the original dissimilarity-based approach [7], in MCD prototypes are representative of *local* components of a given body part, instead of the whole part.

Prototypes are created as follows [5] (see Fig. 1-b, c). For each body part $m = 1, \ldots, M$:

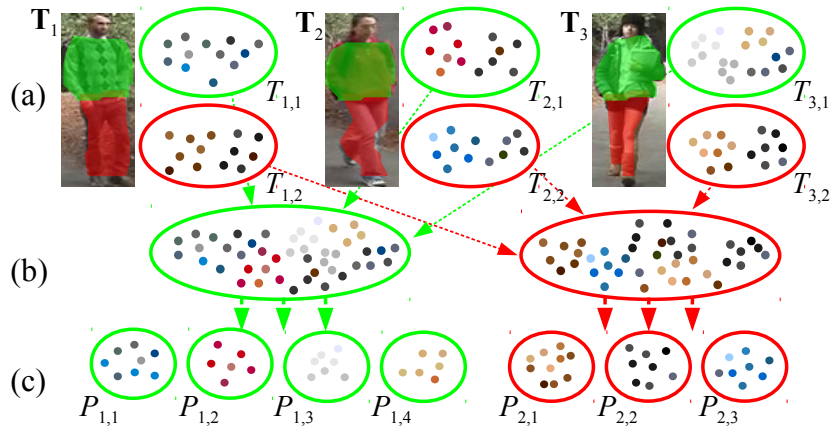


**Fig. 1.** Outline of the MCD framework (taken from [5]). In this example, two body parts are considered, upper (green) and lower (red) body. (a) Representation of three individuals as sets of components, shown as coloured dots. (b-c) Prototype creation: all the components of the same part are (b) merged, and (c) clustered.

1. Merge the feature vectors of the $m$-th part of each $\mathbf{I} \in \mathcal{I}$ into a set $X_m = \bigcup_{j=1}^{N} I_{j,m}$;
2. Cluster the set $X_m$ into a set $\mathbf{P}_m$ of $N_m$ clusters, $\mathbf{P}_m = \{P_{m,1}, \ldots, P_{m,N_m}\}$. Take each cluster as a prototype for the $m$-th body part.

Each prototype is a set of visually similar image components, which can belong to different individuals. In turn, each original descriptor $\mathbf{I}$ consists of a set of components for each body part. Thus, to create a dissimilarity vector from $\mathbf{I}$, dissimilarities can be evaluated via a distance measure between sets. In [5] the *k-th Hausdorff Distance* was used, due to its robustness to outliers.

## 3 A general method for appearance-based people search

We now present a simple and general approach to implement appearance-based people search based on MCD, by extending any re-identification method that uses a multiple part and multiple component representation of clothing appearance. The use of dissimilarity descriptors allows us to define detectors *independently* of the specific body part subdivision and local features used. Previous works required instead the definition of ad-hoc detectors for a given descriptor [3], or focused on a specific kind of descriptor [4]. Our intuition is that the clothing characteristics that can be detected by a given appearance descriptor, according to its low-level features and part subdivision (e.g., "red shirt"), may be encoded by one or more visual prototypes. For example, the rectangular image patches in Fig. 2 are sample components of 10 prototypes, extracted from the upper body parts of individuals taken from the data set of Sect. 4, using the MCD implementation of [5]. Intuitively, descriptors of people wearing a red shirt

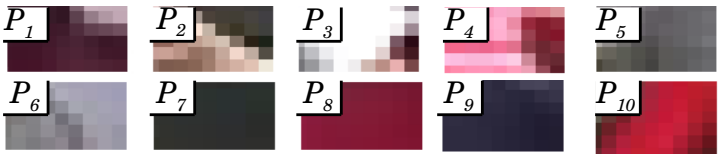


**Fig. 2.** Prototypes obtained from the upper body parts of a set of individuals.

should exhibit a high similarity to prototypes $P_8$ and $P_{10}$, while a high similarity to $P_3$ can be expected in the case of a white shirt.

Following the above intuition, a possible approach to perform appearance-based people search through an existing appearance descriptor, is to: (i) identify a set $\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \ldots\}$ of clothing characteristics that can be detected by the given descriptor, named *basic queries*; (ii) construct a detector for each basic query $\mathbf{Q}_i$, using dissimilarity values as *features* of a supervised classification problem.

The basic queries that have to be identified in step (i) depend on the original descriptor. For instance, if it separates lower and upper body parts, and uses colour features, one basic query can be "red trousers/skirt". Step (ii) can be viewed as a supervised binary classification problem for each $\mathbf{Q}_i$, which consists of recognising the presence or absence of the corresponding visual characteristic, using as features the dissimilarity values between an image descriptor and the prototypes. The training set can be obtained from a gallery of images of individuals, labelled accordingly. The resulting classifier can then be used as the detector for the basic query $\mathbf{Q}_i$. Note that one may know in advance that some features (prototypes) do not carry any discriminant information for some $\mathbf{Q}_i$. For instance, this is the case of the prototypes of the lower body part, with respect to queries related to the upper body. Such features can thus be discarded before constructing the corresponding classifier. Finally, complex queries can be built by connecting basic ones through Boolean operators, e.g., "red shirt AND (blue trousers OR black trousers)". Given a set of images, those relevant to a complex query can simply be found by combining the subsets of images found by each basic detector, using the set operators corresponding to the Boolean ones. In the above example, this amounts to the union (OR) of the images retrieved by the "blue trousers" and "black trousers" basic queries, followed by the intersection (AND) with the images retrieved by the basic query "red shirt".

We point out that the above approach for building detectors is independent of the original appearance descriptor.

## 4 Experimental evaluation

**Implementation.** We evaluated our people search approach using two different descriptors previously proposed for person re-identification. The first descriptor was proposed in [6]. It subdivides body into torso and legs, and represents each part with the HSV colour histograms of a bag of randomly extracted 80 image

patches. The second is the SDALF descriptor proposed in [8]. It uses the same part subdivision above, and represents each part with three local features: an HSV colour histogram, the "Maximally Stable Colour Regions", and the "Recurrent Highly Structured Patches" (RHSP). The first two features are related to the colour, while RHSP codifies the most recurrent repeated patterns. We also used a variation of the first descriptor: it uses a pictorial structure [9] to subdivide body into nine parts: arms and legs (upper and lower, left and right), and torso. The corresponding implementations of our people search method are denoted respectively as $\mathrm{MCD}_1$, $\mathrm{MCD}_2$ and $\mathrm{MCD}_3$.

All the above descriptors enable queries related to clothing colour. $\mathrm{MCD}_1$ and $\mathrm{MCD}_2$ should permit queries related to upper or lower body, like "white upper body garment". $\mathrm{MCD}_3$ should also enable more specific queries, like "short sleeves", that may be distinguished by the presence of skin-like colour in lower arms. Finally, the RHSP feature used in $\mathrm{MCD}_2$ should should enable queries related to textures, like "checked trousers".

Prototypes were obtained by the a two stage clustering scheme as in [5]. In $\mathrm{MCD}_3$, for each body part three different sets of prototypes were created, one for each kind of local features. In the experiments we considered different numbers of prototypes for each body part, ranging from 5 to 300. The $k$-th Hausdorff distance was used to compute dissimilarities, with $k = 10$.

**Data Set.** We used the VIPER data set [10]. It is made up of 1264 images of 632 pedestrians, of size $48 \times 128$ pixels, that exhibit different lighting conditions and pose variations. We defined 14 basic queries related to the colour of the upper and lower body parts, and to the presence of short sleeves/trousers/skirts. They are reported in Table 1, where the corresponding number of relevant images is shown between brackets. We defined these basic queries by considering clothing characteristics that: 1) were detectable to the considered descriptors, and 2) were present in several VIPER images, to allow us to construct a training set of a certain size for building the corresponding descriptors. For constructing the training sets, we needed to manually tag images according to each basic query. We labelled a subset of 512 images, denoted in the following as *VIPER-Tagged*. These images, and the corresponding labels, are available at `http://prag.diee.unica.it/pra/research/reidentification/dataset/viper_tagged`.

**Experimental setup**. We evaluated the retrieval performance of our approach on each basic query, for each considered descriptor, using the precision-recall (P-R) curve. We first extracted the MCD visual prototypes from the whole *VIPER-Tagged* data set. Then, for each basic query we randomly subdivided VIPER-Tagged into a training and a testing sets of equal size, using stratified sampling, and trained a classifier on training images to implement a detector. An SVM classifier with linear kernel was used to this aim. The P-R curve was evaluated on testing images by varying the SVM decision threshold. This procedure was repeated ten times, and the resulting P-R curves were averaged.

We point out that in these experiments we considered an off-line application scenario. In this kind of scenario, the data set in which one want to search is usually entirely available (e.g., in forensic investigations, all the available data

| Class (cardinality) | $MCD_1$ | $MCD_2$ | $MCD_3$ |
|---|---|---|---|
| red shirt (51) | **0.845** | 0.780 | 0.792 |
| blue/light blue shirt (34) | **0.645** | 0.523 | 0.494 |
| pink shirt (35) | 0.534 | **0.578** | 0.461 |
| white/light gray shirt (140) | **0.771** | 0.736 | 0.758 |
| black shirt (156) | 0.728 | 0.705 | **0.736** |
| orange shirt (10) | **0.689** | 0.580 | 0.463 |
| violet shirt (18) | 0.422 | 0.235 | **0.433** |
| green shirt (34) | **0.687** | 0.594 | 0.619 |
| short sleeves (220) | 0.631 | 0.608 | **0.643** |
| red trousers/skirt (16) | 0.713 | 0.638 | **0.916** |
| black trousers/skirt (12) | 0.683 | 0.607 | **0.711** |
| white/light gray trousers/skirt (81) | **0.758** | 0.639 | 0.635 |
| blue/light blue trousers/skirt (175) | **0.641** | 0.622 | 0.620 |
| short trousers/skirt (82) | 0.416 | 0.393 | **0.557** |

**Table 1.** Average break-even point attained using the considered descriptors.

is usually provided to the investigators). In this case, one can conveniently use all the available images for prototype creation, which is an unsupervised procedure, and does not require any manual labelling. In other application scenarios (e.g., on-line), one should instead extract prototypes off-line from a design data set, and use them to compute the dissimilarity representations of newly seen pedestrians at operation phase. In principle, in this case the performance of the proposed method could be lower, if the design set is not representative of the data processed at operation phase. However, the experimental evidences reported in [5] suggest that if a design data set containing a wide range of different clothing characteristics is used for prototype creation, the prototypes should be representative enough for a different set of pedestrians (i.e., those seen at operation phase).

**Results.** The performance on each basic query is summarised in Table 1, in terms of the corresponding average break-even point (BEP), which is the point of the P-R curve whose precision equals recall. The best performance for each basic query is highlighted in bold. In Fig. 3 we report four representative examples of the average P-R curves. An example of the ten top-ranked images for two basic queries is also shown in Fig. 4.

Our method attained a rather good performance with all descriptors, for almost all basic queries. The best performance was attained on basic queries related to the colours red, white and black (see Table 1). The most likely reason is that such colours are well separated in the HSV space, which is used by all the considered descriptors. As pointed out in Sect. 3, $MCD_3$ was likely to attain the best performance on basic queries related to the presence of skin on lower arms and legs, namely "short sleeves" and "short trousers/skirt" (see Fig. 3, bottom-left plot), due to its more refined body subdivision. Nevertheless, also $MCD_1$ and $MCD_2$ attained a good performance on these classes. The reason is that, although $MCD_1$ and $MCD_2$ can not distinguish between lower and upper

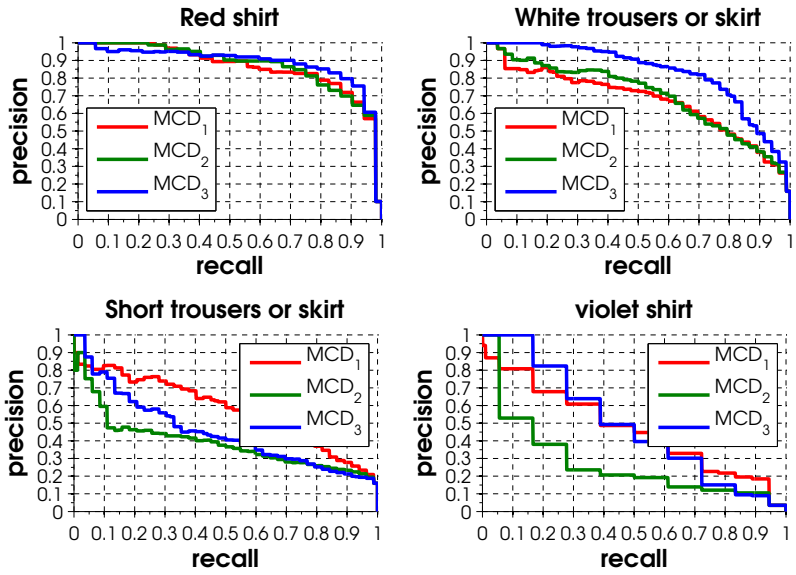


**Fig. 3.** Average P-R curves of 4 queries.

arms (legs), they are nevertheless able to detect skin-like colour in the whole arms or legs.

We finally evaluated how performance is affected by the number of prototypes $N_m$ for each part $m$. We observed that the performance initially grows as $N_m$ increases, then reaches a nearly stable value around $N_m = 100$ (for $\mathrm{MCD}_2$, $\mathrm{MCD}_3$) or 200 (for $\mathrm{MCD}_1$), depending on the basic query. This behaviour can be easily explained: once the number of prototypes is enough so that most of the distinctive visual characteristics have been captured by different clusters, increasing the number of prototypes has mainly the effect of splitting some of the previous clusters into two or more similar ones. Consequently, no further information is embedded in the new prototypes. Note that the results reported in Table 1 and Fig. 3 were attained for $N_m = 200$ (for $\mathrm{MCD}_1$) and $N_m = 100$ (for $\mathrm{MCD}_2$ and $\mathrm{MCD}_3$).

## 5 Conclusions

We proposed a general approach to implement the task of searching images of individuals that match a given *textual* query related to clothing appearance, through the same kind of descriptors used in most existing person re-identification systems, where the query is an *image* of an individual of interest, instead. Our approach is based on turning such descriptors into dissimilarity-based ones, exploiting the framework of [5]. This allows one to add a very useful functionality (e.g., for forensic investigations), to a re-identification system. Our

**Fig. 4.** The top ten images retrieved by MCD1, for the "red shirt" (top) and "short sleeves" (bottom) queries, sorted from left to right for decreasing values of the relevance score provided by the detector (classifier). Note that only one non-relevant image is present, highlighted in red.

approach attained promising results on preliminary experiments with three different descriptors, on a benchmark data set. An interesting direction of further research is to extend our approach to deal with video sequences. To this aim, pedestrian detection and tracking functionalities that should be deployed as part of a person re-identification system, could be exploited. In this case, a bag of dissimilarity vectors coming from different frames would be available for each person, instead of a single one. A Multiple Instance Learning approach [11] could then be used to train the detectors.

## References

1. Doretto, G., Sebastian, T., Tu, P., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. Journal of Ambient Intelligence and Humanized Computing **2** (2011) 127–151
2. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE Transactions on Pattern Anallisys and Machine Intelligence **25** (2003) 1505–1518
3. Vaquero, D., Feris, R., Tran, D., Brown, L., Hampapur, A., Turk, M.: Attribute-based people search in surveillance environments. In: IEEE Workshop on Applications of Computer Vision (WACV'09). (2009)
4. Thornton, J., Baran-Gale, J., Butler, D., Chan, M., Zwahlen, H.: Person attribute search for large-area video surveillance. In: 2011 IEEE International Conference on Technologies for Homeland Security (HST). (2011) 55 –61
5. Satta, R., Fumera, G., Roli, F.: Fast person re-identification based on dissimilarity representations. Pattern Recognition Letters, Special Issue on Novel Pattern Recognition-Based Methods for Reidentification in Biometric Context (2012, in press)
6. Satta, R., Fumera, G., Roli, F., Cristani, M., Murino, V.: A multiple component matching framework for person re-identification. In: Proc. of the 16th Int. Conf. on Image Analysis and Processing (ICIAP). Volume 2. (2011) 140–149

7. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence). World Scientific Publishing Co., Inc., River Edge, NJ, USA (2005)
8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proc. of the 2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2010) 2360–2367
9. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2009) 1014–1021
10. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. of the 10th IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS). (2007) 41–47
11. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89** (1997) 31–71