IET Journals
The Institution of Engineering and Technology

A Multi-Stage Ranking Approach for Fast Person Re-Identification

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Bahram Lavi, Giorgio Fumera ✉*, Fabio Roli*

Department of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123, Cagliari, Italy
✉ *E-mail: fumera@diee.unica.it*

Abstract: One of the goals of person re-identification systems is to support video-surveillance operators and forensic investigators to find an individual of interest in videos acquired by a network of non-overlapping cameras. This is attained by sorting images of previously observed individuals for decreasing values of their similarity with a given probe individual. Existing appearance descriptors, together with their similarity measures, are mostly aimed at improving ranking quality. We address instead the issue of processing time, which is also relevant in practical applications involving interaction with human operators. We show how a trade-off between processing time and ranking quality, *for any given descriptor*, can be achieved through a multi-stage ranking approach inspired by multi-stage classification approaches, which we adapt to the re-identification ranking task. We analytically model the processing time of multi-stage system and discuss the corresponding accuracy, and derive from these results practical design guidelines. We then emprically evaluate our approach on three benchmark data sets and four state-of-the-art descriptors.

1 Introduction

Person re-identification is the computer vision task of recognizing an individual over a network of video surveillance cameras with non-overlapping fields of view [1]. One of its applications is to support surveillance operators and forensic investigators in retrieving videos showing an individual of interest, given an image as a query (*probe*). To this aim, the video frames or tracks of all the individuals (*template gallery*) recorded by the camera network are ranked in order of decreasing similarity to the probe, to allow the user to find the occurrences (if any) of the individual of interest hopefully in the top positions. This is a challenging task in typically surveillance videos, due to low image resolution, unconstrained pose, illumination changes, and occlusions, which do not allow to exploit strong biometrics like face. Clothing appearance is therefore the most widely used cue; other cues like gait and anthropometric measures have also been investigated. Most of the existing techniques are based on defining a specific descriptor of clothing appearance (typically including color and texture), and a specific similarity measure between a pair of descriptors (evaluated as a *matching score*) which can be either manually defined or learnt from data [1–5].

The main focus of existing work in this field is to attain a high ranking accuracy. Processing time is an issue which has received much less attention so far, instead (to our knowledge, only in [6–8]), despite its relevance in practical applications involving interaction with human operators, like the ones mentioned above. Many of the existing similarity measures (either hand-crafted or learnt from data) are indeed rather complex, and require a relatively high processing time, e.g., [3, 5, 9, 10]. On the other hand, in real-world applications the template gallery can be very large, and even if the processing time for a single matching score is low (e.g., the Euclidean distance between fixed-length feature vectors [5]), evaluating the matching scores for all the templates can be time-consuming.

One possible solution to reduce processing time is to reduce the complexity of a given descriptor and/or of the associated similarity measure; however, this is likely to reduce ranking accuracy as well. A known approach in the pattern recognition field, in particular for supervised classification systems, to trade a lower classification accuracy for a lower processing time, is to use a multi-stage architecture (e.g., [11, 12]). Inspired by this approach, in this paper we investigate whether and how a multi-stage architecture can be exploited to attain an analogous trade-off between ranking accuracy also in person re-identification systems. In particular, we focus on attaining such a trade-off for any single, *given* descriptor, with no constraint on the descriptor itself. Since existing multi-stage solutions cannot be directly applied to person re-identification, which involves a *ranking* problem rather than a *classification* one, we first provide a formalization of multi-stage ranking systems: we develop an analytical model of their processing time, and discuss the behaviour of the corresponding ranking accuracy. Based on these results we propose practical design criteria for multi-stage person re-identification systems, considering applications requirements given in terms of a constraint on the maximum allowed processing time.

The main contribution of this work is the extension of the multi-stage architecture used in pattern classification to person re-identification (using *any* given descriptor and similarity measure), by formalizing the underlying multi-stage ranking problem and by studying the resulting accuracy-time trade-off; this also allow us to suggest practical design criteria. This work extends our preliminary work [13] in the analytical model of the behaviour of multi-stage re-identification systems, in the design criteria, and in a wider empirical investigation.

This paper is structured as follows. We first summarize related work in Sect. 2. In Sect. 3 we formalize multi-stage ranking problems and develop design criteria for multi-stage person re-identification systems. In Sect. 4 we evaluate their effectiveness on three benchmark data sets, using four state-of-the-art descriptors.

2 Related work

In this section we first describe existing multi-stage approaches to classification problems. We then summarize person re-identification techniques aimed at reducing processing time in the computation of matching scores, and/or based on a multi-stage ranking approach.

2.1 Multi-stage classification approaches

The multi-stage approach is used since a long time in pattern classification systems. For instance, in [14] a cascade of classifiers was proposed to attain a trade-off between classification accuracy and the cost of feature acquisition, e.g., for medical diagnostics applications: each classifier uses features that are more discriminant, but also more costly [14] than previous classifiers. The goal is to assign

an input instance (e.g., a medical image) to one of the classes (e.g., the outcome of a diagnosis) with a predefined level of confidence, using features (e.g., medical exams) with the lowest possible cost; if a classifier but the last one does not reach the desired confidence level, it *rejects* the input instance (i.e., withholds making a decision), and sends it to the next stage. This approach has later been exploited to attain a trade-off between classification accuracy and processing time, e.g., in handwritten digit classification [11, 15, 16]. A similar approach is used in the well-known algorithm of [12] for designing fast object detectors: it consists in detecting and discarding background regions of the input image as quickly as possible, using classifiers based on features fast to compute; this allows focusing the attention on regions more likely to contain the object of interest, using classifiers based on more discriminative features that also require a higher processing time.

2.2 Reducing processing time in person re-identification

To our knowledge, the issue of processing time has been explicitly addressed so far in the context of person re-identification only in [6–8]. Only in [7] the proposed solution is a multi-stage system: the first stage selects a subset of templates using a descriptor which is built upon a bag-of-words feature representation and an indexing scheme based on inverted lists, and requires a low processing time for computing matching scores; the second stage ranks only the selected templates using a different, more complex descriptor based on mean Riemann covariance. Differently to our approach, in [7] only two stages are considered, and only a subset of templates is ranked by the whole system, possibly losing the correct identity. Moreover, a different, specific descriptor is used in each stage, whereas our approach can be applied to any descriptor, and uses different versions of the *same* descriptor at each stage. In [6] we proposed a dissimilarity-based approach to design descriptors made up of bags of local features, possibly extracted from different body parts. It consists in finding a set of M representative local features (called prototypes) from all individuals of the template gallery, and in representing each template and probe image as a vector of M dissimilarity values between the corresponding bag of local features and the templates. This allows the matching score to be computed as a distance between feature vectors, rather than using a more complex similarity measure between bags of local features. Contrary to the multi-stage approach proposed in this paper, the one of [6] can be applied to descriptors made up of bags of local features. The method of [8] reduces processing time in the specific multi-shot setting (when several images per individual are available), and for specific descriptors based on local feature matching, e.g., interest points. It first filters out irrelevant interest points, then it builds a sparse representation of the remaining ones.

2.3 Multi-stage re-identification systems

Multi-stage re-identification systems have already been proposed by some authors. Their aim is however to improve ranking accuracy, without taking into account processing time [17–21]. In [17] the first stage uses returns the operator the 50 top-ranked templates; if the probe identity is not among them, a classifier is trained to discriminate the probe image from other identities, and is used to re-rank the remaining templates. In [18] person re-identification is addressed as a content-based image retrieval task with relevance feedback, for settings where several instances of a probe can be present in the template gallery; accordingly, the aim is to increase recall. In each stage (i.e., iteration of relevance feedback) only the top-ranked templates are shown to the operator, then his feedback is exploited to adapt the similarity measure for the probe at hand, and the remaining templates are re-ranked. A similar multi-stage strategy was proposed in [19] for reducing the operator's effort in analyzing the template images: in each stage only the top-ranked templates are presented to the operator, who is asked to select a "strong negative" (i.e., a different individual whose appearance is most dissimilar to the probe), and optionally a few "weak negatives"; a post-rank function is then learnt based on this feedback and on the probe image, and the remaining templates are re-ranked in the next stage. A similar, two-stage approach was proposed in [20]: the operator is asked to label some pairs of locally similar and dissimilar horizontal image regions in the top-ranked templates, and this feedback is exploited to re-rank all templates. Another two-stage approach was proposed in [21], to improve the ranking provided by a given first-stage descriptor: a small subset of the top-ranked templates is re-ranked by the second stage, by a different descriptor that uses a manifold-based method with three specific low-level features.

3 Multi-stage re-identification systems

As pointed out in Sect. 1, existing multi-stage approaches to classification problems, aimed at trading classification accuracy for the processing time, cannot be directly applied to person re-identification, which involves a ranking problem. As the main contribution of this work, in this section we develop a specific formulation of multi-stage ranking problems focused on trading ranking accuracy for processing time in person re-identification systems, for any given descriptor and similarity measure. In particular, we first develop an analytical model of processing time and discuss the behaviour of the corresponding ranking accuracy measured using the CMC curve. Based on these results we also propose practical design criteria.

3.1 Problem definition and notation

Let D denote a given descriptor, $m(\cdot,\cdot)$ the corresponding similarity measure between a pair of images, t the processing time for computing it, $\mathbf{T}$ and $\mathbf{P}$ the descriptors of a template and probe image, respectively, and $G = \{\mathbf{T}_1, \ldots, \mathbf{T}_n\}$ the template gallery. For a given probe $\mathbf{P}$, a standard re-identification system computes the matching scores $m(\mathbf{P}, \mathbf{T}_i)$, $i = 1, \ldots, n$, and returns the list of template images ranked in order of decreasing values of their score. Ranking accuracy is typically evaluated using the CMC curve, defined as the probability that the correct identity is within the first r ranks, for $r = 1, \ldots, n$. By definition, the CMC curve increases with r, and equals 1 for $r = n$.

In this paper we consider application scenarios characterized by strict requirements on the processing time for obtaining the ranked list of templates, e.g., due to real-time constraints. In particular, we consider requirements expressed by the constraint $t \leq t_{\max}$, where $t_{\max}$ is an application-specific value. Many existing appearance descriptors attain a high recognition rate at the expense of a high complexity, which results in a relatively high value of t, e.g., [3, 5, 9, 10]. Moreover, even if t is relatively low, when the gallery set size is very large an even lower $t_{\max}$ value may be required. Focusing on the case when a *given* descriptor D exhibits a satisfactory ranking accuracy, but does not meet the constraint $t \leq t_{\max}$, in the next section we propose a multi-stage ranking approach capable of trading a lower ranking accuracy for a lower processing time.

3.2 A multi-stage ranking approach for person re-identification

Let us first discuss the case of a two-stage ranking system. Consider a given descriptor, that we denote as D_2, and assume that it exhibits a satisfactory ranking accuracy (CMC curve) but a too high processing time, $t_2 > t_{\max}$, as explained above. Our approach is based on modifying D_2, by changing its parameters, into a descriptor D_1 that exhibits a lower processing time $t_1 < t_{\max}$. Usually this can be attained only at the expense of a lower accuracy, i.e., the CMC curve of D_1 (denoted as CMC_1) lies below that of D_2 (CMC_2). If CMC_1 is not satisfactory for the application at hand, D_1 and D_2 can be combined into a two-stage system to meet the constraint on processing time, attaining at the same time a CMC curve better than CMC_1. To this aim, for a given probe, first all n templates are ranked using D_1, then the n_2 top-ranked ones are re-ranked using D_2, for a given n_2, with $1 < n_2 < n$. The resulting

average processing time per probe, t_{1-2}, is given by:

$$t_{1-2} = \frac{1}{n} t_{\mathrm{D}_1} + t_1 + \frac{n_2}{n} t_2 \,, \tag{1}$$

where also the the time t_{D_1} for computing the descriptor D_1 of the probe is taken into account (the same descriptor can be computed offline for templates, and is therefore not considered). Note that the impact of such an overhead time reduces as the overall number of templates to be ranked increases. From Eq. (1), the constraint $t_{1-2} \leq t_{\max}$ translates into:

$$n_2 \leq n \frac{(t_{\max} - t_1)}{t_2} - \frac{t_{\mathrm{D}_1}}{t_2} \,. \tag{2}$$

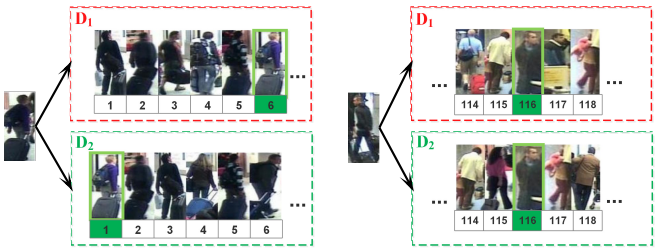

Fig. 1: Two examples of the ranked list of templates produced by a descriptor D_2 and by a less accurate version of it, D_1, for a given probe (the correct identity is marked in green). Left: the correct identity is in the top ranks, and is ranked *higher* by D_2. Right: the correct identity has a low rank, and is ranked *identically* by both descriptors.

Consider now the resulting CMC curve, denoted by CMC_{1-2}. To make an analytical derivation of its behaviour possible, at least to some extent, we disregard the general case when CMC_1 and CMC_2 cross in one or more points, and consider only the case when $CMC_1(r) < CMC_2(r)$ for ranks $r \leq r^*$, and $CMC_1(r) = CMC_2(r)$ for $r > r^*$, for a given rank $r^* \leq n$, as in the example of Fig. 3. In other words, when D_2 gives a rank between 1 and r^* to the template of the correct identity, the rank given by D_1 is on average lower; when D_2 gives a rank between r^* and n, instead, the rank given by D_1 is on average the same (see the example in Fig. 1). In the limit cases of $n_2 = 1$ and $n_2 = n$, it is easy to see that $CMC_{1-2} = CMC_1$ and $CMC_{1-2} = CMC_2$, respectively. For $1 < n_2 < n$, the above assumption implies that CMC_{1-2} lies between CMC_1 and CMC_2, and approaches CMC_2 as n_2 increases. This can be proven as follows. First, $CMC_{1-2}(r) = CMC_1(r)$ for all $r \geq n_2$, since for any $r \geq n_2$ the correct identity is among the r top ranks of the two-stage system, if and only if it is among the r top ranks of D_1. Second, since the n_2 top-ranked templates by D_1 are re-ranked by the more accurate D_2, it follows that $CMC_{1-2}(r) \geq CMC_1(r)$ for $r < n_2$. An example of this behaviour is reported in Fig. 3 for two different values of n_2.

To sum up, for two-stage systems a trade-off between processing time and ranking accuracy can be attained by values of n_2 that satisfy constraint (2): the higher n_2, the higher the resulting processing time and ranking accuracy.

The above results can be generalized to multi-stage systems with $N > 2$, using the original descriptor in the last stage as D_N, and different versions of D in the previous stages as $\mathrm{D}_1, \ldots, \mathrm{D}_{N-1}$, characterized by increasing ranking accuracy and increasing processing time, $t_1 < t_2 < \ldots < t_{\mathrm{D}_N}$, with $t_1 < t_{\max}$ (see Fig. 2). Denoting by n_i the number of matching scores computed by the i-th stage, under the constraint:

$$n_1 = n > n_2 > \ldots > n_N > 1 \,, \tag{3}$$

the corresponding average processing time $t_{1-\mathrm{N}}$ is:

$$t_{1-\mathrm{N}} = \frac{1}{n} \sum_{i=1}^{N-1} t_{\mathrm{D}_i} + t_1 + \sum_{i=2}^{N} \frac{n_i}{n} t_i \,. \tag{4}$$

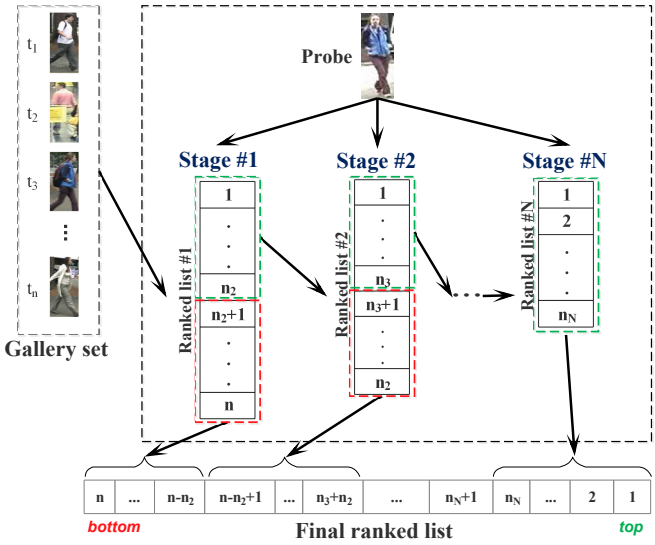

Fig. 2: Scheme of the proposed multi-stage ranking approach.

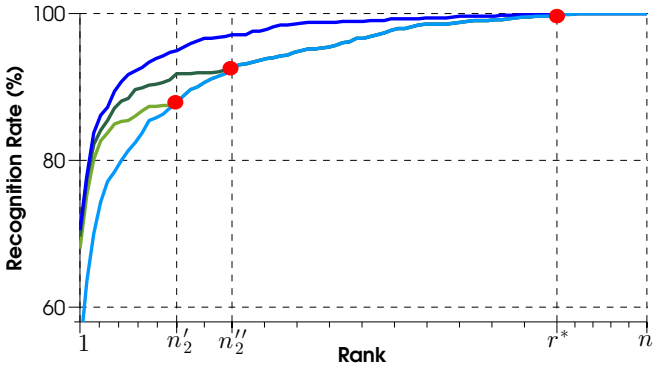

Fig. 3: Example of CMC curves of two-stage systems. Light blue: first-stage; dark blue: second-stage; r^* is the rank from which their CMC curves become identical; light and dark green: two-stage systems corresponding to different values of n_2.

Accordingly, the constraint $t_{1-\mathrm{N}} \leq t_{\max}$ can be rewritten as:

$$\sum_{i=2}^{N} n_i t_i \leq n(t_{\max} - t_1) - \sum_{i=1}^{N-1} t_{\mathrm{D}_i} \,. \tag{5}$$

Note that constraint (2) is a particular case of (5) for $N = 2$.

Assuming that the CMC curves of any pair of adjacent stages, CMC_i and CMC_{i+1}, exhibit the same behaviour considered above (see Fig. 3), by the same arguments above it follows that the CMC curve of the multi-stage system, CMC_{1-N}, lies between CMC_1 and CMC_N. In particular, in the limit cases when $n_i = 1$, and $n_i = n$ for every $i > 1$, we obtain $CMC_{1-N} = CMC_1$, and $CMC_{1-N} = CMC_N$, respectively. Moreover, $CMC_{1-N}(r) = CMC_1(r)$ for $r \geq n_2$. In general, for increasing values of $n_2, \ldots, n_N$, CMC_{1-N} gets closer to CMC_N.

Accordingly, for a generic multi-stage system a trade-off between processing time and accuracy can be attained when the n_i's satisfy constraints (3) and (5); the higher $n_2, \ldots, n_N$, the higher the resulting processing time and ranking accuracy.

3.3 Design criteria

Designing a multi-stage re-identification system according to the above approach requires to choose the number N of stages, the descriptors $\mathrm{D}_1, \ldots, \mathrm{D}_{N-1}$, and the number of templates $n_2 > \ldots > n_N$ to be re-ranked at each stage, under constraints (3) and (5). The

best solution, among the ones that satisfy such constraints, is the one that maximizes ranking accuracy. However it cannot be analytically found, and finding it empirically by evaluating all possible choices is clearly impractical, since the three choices above are interrelated and many possible solutions may even exist. In the following we discuss each choice separately, and suggest practical, though suboptimal, design criteria.

Descriptors. Consider first the problem of developing different versions D_{N-1}, ..., D_1 of a given descriptor D, exhibiting a decreasing ranking accuracy and a decreasing processing time. This can be attained by suitably modifying the parameters of D. However existing descriptors can be very complex and contain several parameters. Moreover, only an empirical evaluation is usually possible of the impact of any parameter on ranking accuracy; for instance, the relative behaviour of the CMC curves of any two descriptors depends on the data at hand: see, e.g., the CMC curves of the original and of the first-stage SDALF descriptor on the VIPeR and ETHZ1 data sets, in Fig. 4. To define a practical design criterion we propose to subdivide descriptors into two main categories: fixed-size feature vectors (e.g., [5, 10]), and descriptor with variable size (e.g., [3]). For fixed-size feature vectors, an unsupervised feature reduction technique like PCA can be used. The suitability of PCA to person re-identification tasks is witnessed to its use in the pre-processing step of gBiCov [5]. For descriptors with variable size we suggest to modify the parameter that has the highest impact on processing time; for instance, in SDALF descriptor [3] such a parameter is the number of "blobs" of its MSCR component (see Sect. 4.1). Once a *single* parameter has been chosen (either the feature set size for the former category of descriptor, or a descriptor-specific parameter for the latter category), its value for each stage (but the last one) can be set according to the corresponding processing time, with the only constraint is that the left-hand side of inequality (5) is positive, which amounts to:

$$t_1 < t_{\max} - \frac{1}{n} \sum_{i=1}^{N-1} t_{\mathrm{D_i}} \, . \tag{6}$$

As a simple guideline, one should set t_1 to be no more than half the above upper bound. In particular, if a descriptor-specific parameter is modified, the resulting processing time may need to be empirically evaluated. Instead, if the feature set size is changed, the resulting reduction in processing time can be simply considered as proportional to the reduction in feature set size for common distance measures used for fixed-size feature vectors, like Euclidean and cosine distance.

Number of templates to be re-ranked at each stage. Assuming that N and D_{N-1}, ..., D_1 have already been chosen, the choice of $n_2, \ldots, n_N$ can be discussed separately for $N = 2$ and $N > 2$. For two-stage systems, the single value of n_2 has to be chosen under constraint (2). In this case the best trade-off between processing time and ranking accuracy can be identified a priori: it is attained when the second stage re-ranks the highest possible number of templates, which leads to:

$$n_2 = \left\lfloor n \frac{(t_{\max} - t_1)}{t_2} - \frac{t_{\mathrm{D_1}}}{t_2} \right\rfloor \, . \tag{7}$$

In the case $N > 2$, constraints (3) and (5) define a convex polyhedron in the $N-1$-dimensional space, and the feasible solutions are all the points $\mathbf{n} = (n_2, \ldots, n_N)$ with integer coordinates belonging to such a polyhedron. However, among these solutions it is not possible to identify a priori the one that maximizes ranking accuracy. One can only discard the *dominated* solutions: if a solution $\mathbf{n}' = (n'_2, \ldots, n'_N)$ is dominated by a different solution $\mathbf{n}'' = (n''_2, \ldots, n''_N)$, i.e., $n'_2 \leq n''_2, \ldots, n'_N \leq n''_N$, then $\mathbf{n}'$ can be discarded, since each of its stages (but the first one) re-ranks a lower or identical number of templates than the corresponding stage of $\mathbf{n}''$, and consequently its ranking accuracy will be lower. Instead, for any pair of non-dominated solutions $\mathbf{n}'$ and $\mathbf{n}''$, if $n'_i < n''_i$ for some i, then some j exists such that $n'_j > n''_j$; this means that their relative ranking accuracy can be evaluated only empirically, which is impractical if the number of non-dominated solutions is high.

To avoid such problems, we consider a simpler, though potentially suboptimal criterion for multi-stage systems with $N > 2$: we consider values of $n_2, \ldots, n_N$ such that, beside satisfying constraints (3) and (5), the number of templates between two consecutive stages is reduced by a same amount $\alpha < 1$, i.e.:

$$n_i = \lfloor \alpha n_{i-1} \rfloor, \; i = 2, \ldots, N \, . \tag{8}$$

It is now easy to see that ranking accuracy is maximized by choosing the maximum value of α that satisfies constraints (3) and (5), which can be found by a simple line search.

Number of stages. Taking into account the design criteria suggested above, we suggest to limit the choice of the number of stages to two or three, to avoid a time-consuming empirical evaluation of more alternatives. In practice, for a two-stage system one can set the parameter of D_1 such that $t_1 < \frac{1}{2}\left(t_{\max} - \frac{1}{n} t_{\mathrm{D}_1}\right)$ (see above); for a three-stage system one can set the parameter of D_1 and D_2 such that $t_1 < \frac{1}{2}\left[t_{\max} - \frac{1}{n}\left(t_{\mathrm{D}_1} + t_{\mathrm{D}_2}\right)\right]$, and t_2 about twice t_1. Then the choice between a two- and a three-stage system can be made based on an empirical comparison of the corresponding ranking accuracy.

4 Experimental evaluation

We evaluated our approach on three benchmark data sets and four state-of-the-art appearance descriptors. We designed two- and three-stage systems as suggested by our design criteria.*

We used the VIPeR, i-LIDS and ETHZ data sets. VIPeR [2] is a challenging dataset for person re-identification; it contains two images of 632 individuals from two camera views, with pose and illumination changes, cropped and scaled to 128×48 pixels. i-LIDS contains 476 images of 119 pedestrians taken at an airport hall from non-overlapping cameras, with pose and lightning variations and strong occlusions. ETHZ contains three video sequences of a crowded street from two moving cameras; images differ in size and exhibit illumination changes, scale variations, and occlusions. We used only the first sequence "SEQ. #1" (ETHZ1) which contains the largest number of pedestrians (83), and 4,857 images in total. We rescaled the images of i-LIDS and ETHZ1 to the same size of 128×48 pixels as in VIPeR, to get a similar processing time.

4.1 Descriptors

We used the SDALF, gBiCov, LOMO and MCM descriptors. gBiCov and LOMO are fixed-size descriptors: according to our suggested design criteria we obtained faster and less accurate versions of each of them by using PCA. SDALF and MCM have not a fixed size, instead: we chose ad hoc parameters to modify as described below. Since our aim was not to fine-tune these descriptors to maximize their performance on each data set, we chose the parameter values by preliminary experiments, and used the same versions of each descriptor for all data sets.

SDALF† [3] subdivides body into four parts: left and right, torso and legs. Three kinds of features are extracted from each part: maximally stable color regions (MSCR), i.e., elliptical regions (blobs) exhibiting distinct color patterns (their number depends on the specific image), with a minimum size of 15 pixels; a $16 \times 16 \times 4$-bins weighted HSV color histogram (wHSV); and recurrent high-structured patches (RHSP) that characterize texture. A specific similarity measure is defined for each feature; the matching score is computed as their linear combination. In our experiments we did not use RHSP, due to its relatively lower performance. The parameter that mostly affects the processing time for computing the matching score turns out to be the blob size used in MSCR. We obtained faster and less accurate versions of SDALF by increasing the minimum MSCR blob size to 65 and to 45 for the first and second stage,

*The source code of our experiments is available at `https://github.com/bahramlavi/MultiStageRanking`

†Source code: `http://www.lorisbazzani.info/sdalf.html`

Table 1 Average processing time t_i (in sec.) for computing one matching score in the i-th stage, for each of the four descriptors. Note that the original descriptor is used in the last stage.

		SDALF	*gBiCov*	*LOMO*	*MCM*
two-stage system	t_1	2.08	0.0003	0.0008	0.060
three-stage system	t_1	1.60	0.0002	0.0003	0.051
	t_2	2.08	0.0003	0.0008	0.060
last stage	t	9.44	0.040	0.037	27.40

respectively (which reduces the number of blobs). We also modified the wHSV histogram by reducing the corresponding number of bins of to $3 \times 3 \times 2$ and to $8 \times 8 \times 3$, given that this is a very easy parameter to change.

gBiCov* [5] is based on biologically-inspired features (BIF) obtained by Gabor filters with different scales over the HSV color channels. The resulting images are subdivided into overlapping regions of 16×16 pixels; each region is represented by a covariance descriptor that encodes shape, location and color information. BIF and covariance descriptors are concatenated, and PCA is used to reduce its dimension to a predefined value. We obtained different versions of gBiCov by reducing its dimension to 5 for two-stage systems, and to 2 and 5 for three-stage systems.

LOMO† [10] extracts an $8 \times 8 \times 8$-bins HSV histogram and two scales of the Scale Invariant Local Ternary Pattern histogram (characterizing texture) from overlapping windows of 10×10 pixels; it then retains one only histogram from all windows at the same horizontal location, obtained as the maximum value among all the corresponding bins. These histograms are concatenated with the ones computed on a down-sampled image. A metric learning method is used to define the similarity measure. We used PCA to reduce the dimension of the LOMO descriptor to 20 for two-stage systems, and to 5 and 20 for three-stage systems.

MCM‡ [9] subdivides body into torso and legs, and extracts 80 randomly positioned image patches from each part. Each patch is represented by a $24 \times 12 \times 4$-bins HSV histogram. Artificial patches are also generated to improve robustness to illumination changes, by changing the brightness and contrast of the original patches in the RGB color channel. The similarity measure is the average k-th Hausdorff distance between the set of patches of each pair of corresponding body parts, where k was set to 10 in [9]. The number of patches is the parameter that mostly affects processing time for computing similarity scores. We obtained different versions of MCM by reducing the number of patches to 10 and to 20 for the first and second stage, respectively. Similarly to SDALF, we also reduced the corresponding number of bins of the HSV histogram to $3 \times 3 \times 2$ and $12 \times 6 \times 2$.

4.2 Experimental setup

For each descriptor D we designed two- and three-stage systems; for the sake of simplicity we used the same version of D to implement D_1 in two-stage and D_2 in three-stage systems. As in [3], for each data set we repeated our experiments on ten different subsets of individuals, using one image of each individual as template and one as probe, and reported the average CMC curve over the ten runs. We used an Intel Core i5 2.6 GHz CPU. We considered three different values of $t_{\max}$ defined as a fraction of the processing time of the original descriptor used in the last stage, $t_{\max} = \beta t_N$, for $\beta = 0.3, 0.4, 0.5$.

Table 2 Number of templates processed at each stage for each descriptor and data set, and for the different values of β.

Descriptor	Data set	Two-stage systems				Three-stage systems					
			β=0.3	β=0.4	β=0.5	β=0.3		β=0.4		β=0.5	
		n	n_2	n_2	n_2	n_2	n_3	n_2	n_3	n_2	n_3
SDALF	VIPeR	316	25	57	88	84	22	120	45	150	71
	i-LIDS	119	9	21	33	31	8	45	17	56	26
	ETHZ1	83	7	15	23	22	5	31	11	39	18
gBiCov	VIPeR	316	92	124	156	170	91	197	123	221	154
	i-LIDS	119	35	47	59	64	34	74	46	83	58
	ETHZ1	83	24	33	41	44	23	51	31	58	40
LOMO	VIPeR	316	88	120	151	167	88	194	119	218	150
	i-LIDS	119	33	45	57	63	33	73	44	82	56
	ETHZ1	83	23	31	40	43	22	51	31	57	39
MCM	VIPeR	316	94	126	157	172	93	199	125	222	156
	i-LIDS	119	35	47	59	64	34	74	46	83	58
	ETHZ1	83	25	33	41	45	24	52	32	58	40

4.3 Results

The average processing time for computing one matching score at each stage, evaluated on VIPeR, is reported in Table 1. Similar processing times were observed in the other data sets, due to the use of the same image size. Note that processing time of MCM cannot be compared to the one of the other descriptors, since MCM was implemented in C# and the other descriptors in Matlab. Note also that the original MCM descriptor has a much higher processing time than its versions used in the first and (for three-stage systems) second stage, with respect to the other descriptors: this is due to the use of the Hausdorff distance as similarity measure, which makes the processing time proportional to the *square* of the number of patches (see Sect. 4.1).

The number of templates processed at each stage, chosen according to the proposed design criterion, is reported in Table 2. The average CMC curves are shown in Figs. 4 and 5, respectively for two- and three-stage systems.

Note first that, since we did not fine-tune the different versions of each descriptor to each data set, in some cases the first and last stages turned out to exhibit very similar CMC curves, and therefore the CMC curve of the corresponding multi-stage systems is similar to both. For instance, this is the case of SDALF and MCM on VIPER, in two-stage systems (Fig. 4). Also in this cases the processing time is nevertheless lower than the one of the original descriptors.

In all the other cases the trade-off between the ranking accuracy and the processing time (given by $t_{\max}$) of multi-stage systems clearly emerges; see, e.g., the CMC curves of SDALF on ETHZ1, both in two- and in three-stage systems. In particular, note that in the top ranks the CMC curve of these multi-stage systems is almost identical to the one of the corresponding original descriptor; it then decreases, starting from a rank that depends on the specific data set and descriptor, up to becoming identical since rank n_2 to the CMC curve of the first stage. Moreover, for a given data set and descriptor, the CMC curve of the corresponding multi-stage system worsens as $t_{\max}$ decreases, i.e., as n_2 (and, for three-stage systems, n_3) increases. We point out that this behaviour agrees with the one that we derived analytically in Sect. 3.2, and then exploited in Sect. 3.3 to define the proposed design criterion. The above results provide evidence that multi-stage re-identification systems designed according to our approach can attain an effective trade-off between processing time and ranking accuracy.

*Source code: `http://vipl.ict.ac.cn/members/bpma`

†`http://www.cbsr.ia.ac.cn/users/scliao/projects/lomo_xqda/`

‡*The source code is available upon request to the authors.*

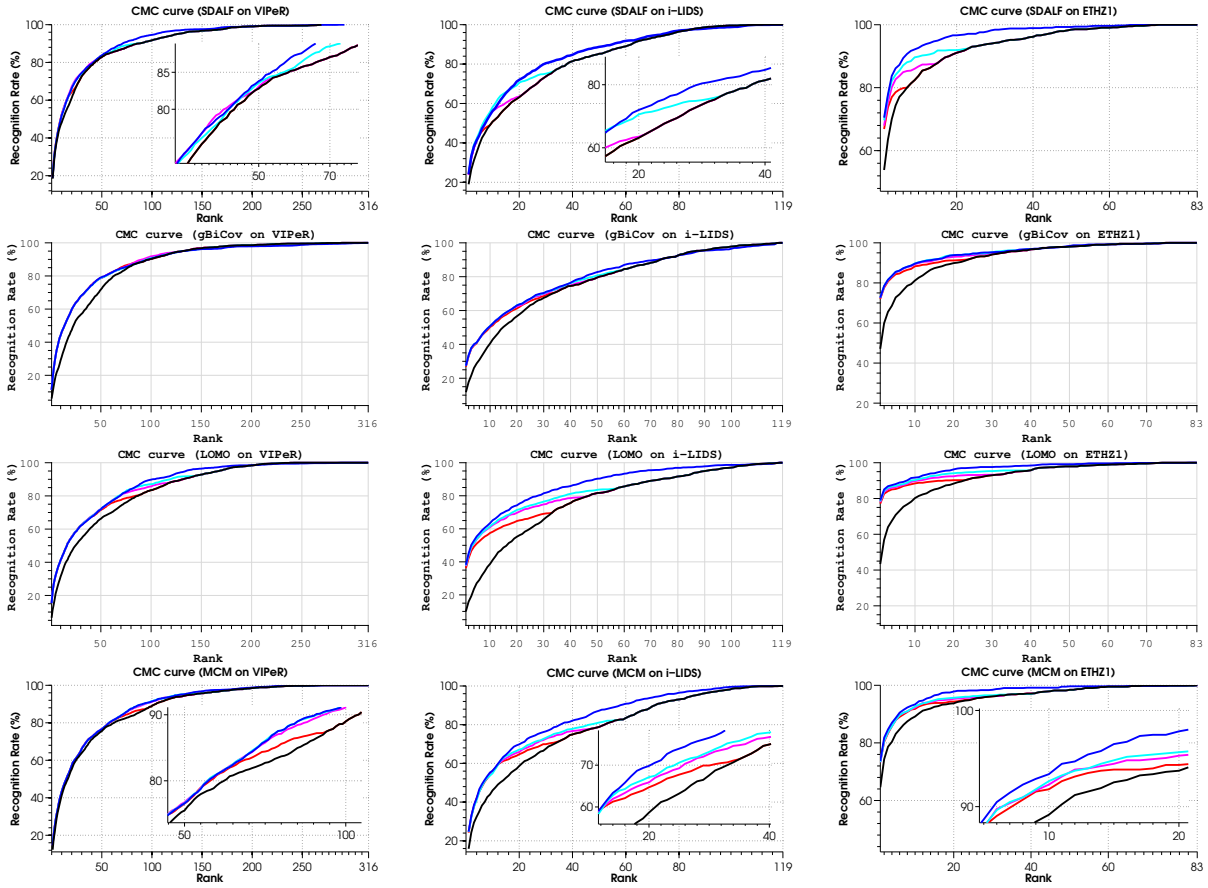

Fig. 4: CMC curves of two-stage systems. Black: first stage; blue: second stage (original descriptor); red, pink, and cyan: two-stage systems with $\beta = 0.3, 0.4, 0.5$, respectively. Enlarged version of plots with very close CMC curves are shown for better visualization.

4.4 Data sets

5 Conclusions

We proposed a multi-stage ranking approach for person re-identification, aimed at trading a lower processing time for a lower ranking accuracy for any *given* appearance descriptor. Our approach is inspired by the well-known multi-stage classification architecture used in pattern recognition systems, which we adapted to ranking problems by developing an ad hoc analytical model of the trade-off between their ranking accuracy and processing time. We also suggested practical design criteria based on our analytical model, and carried out a first empirical investigation on benchmark data sets and state-of-the-art descriptors. Multi-stage re-identification systems can be useful in practical applications that involve interaction with human operators and are characterized by very large template galleries and/or complex descriptors, requiring strict constraints on processing time. They can be useful also in application scenarios when the operator cannot or does not want to scan all the ranked template images (e.g., in real-time settings): in this case, only the subset of templates ranked by the last stage can be returned to the operator. If needed, the attainable trade-off between processing time and ranking accuracy can be improved, with respect to our suggested design criteria, by fine-tuning the different system parameters discussed in Sect. 3.3, at the expense of an additional effort to empirically evaluate the different alternatives.

Acknowledgment

This work has been partially supported by the LETSCROWD project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 740466.

6 References

1 Bedagkar.Gala, A., Shah, S.K.: 'A survey of approaches and trends in person re-identification', *Image and Vision Computing*, 2014, **32**, (4), pp. 270–286
2 Gray, D., Tao, H. 'Viewpoint invariant pedestrian recognition with an ensemble of localized features'. In: Computer Vision–ECCV 2008. (Springer, 2008. pp. 262–275
3 Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M. 'Person re-identification by symmetry-driven accumulation of local features'. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. (IEEE, 2010. pp. 2360–2367
4 Hirzer, M., Roth, P.M., Bischof, H. 'Person re-identification by efficient impostor-based metric learning'. In: Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on. (IEEE, 2012. pp. 203–208
5 Ma, B., Su, Y., Jurie, F.: 'Covariance descriptor based on bio-inspired features for person re-identification and face verification', *Image and Vision Computing*, 2014, **32**, (6), pp. 379–390
6 Satta, R., Fumera, G., Roli, F.: 'Fast person re-identification based on dissimilarity representations', *Pattern Recognition Letters*, 2012, **33**, (14), pp. 1838–1848
7 Dutra, C.R., Schwartz, W.R., Souza, T., Alves, R., Oliveira, L. 'Re-identifying people based on indexing structure and manifold appearance modeling'. In: Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on. (IEEE, 2013. pp. 218–225
8 Khedher, M.I., El.Yacoubi, M.A. 'Two-stage filtering scheme for sparse representation based interest point matching for person re-identification'. In: Advanced Concepts for Intelligent Vision Systems. (Springer, 2015. pp. 345–356
9 Satta, R., Fumera, G., Roli, F., Cristani, M., Murino, V. 'A multiple component matching framework for person re-identification'. In: Image Analysis and Processing–ICIAP 2011. (Springer, 2011. pp. 140–149
10 Liao, S., Hu, Y., Zhu, X., Li, S.Z. 'Person re-identification by local maximal occurrence representation and metric learning'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (, 2015. pp. 2197–2206
11 Sperduti, A.: 'Theoretical and experimental analysis of a two-stage system for classification', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2002, **24**, (7), pp. 893–904
12 Viola, P., Jones, M.: 'Rapid object detection using a boosted cascade of simple features', *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2001, **1**, pp. 511
13 Lavi, B., Fumera, G., Roli, F. 'A multi-stage approach for fast person re-identification'. In: Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+SSPR 2016, Mérida, Mexico, November 29 - December 2, 2016, Proceedings. (Springer International Publishing, 2016.

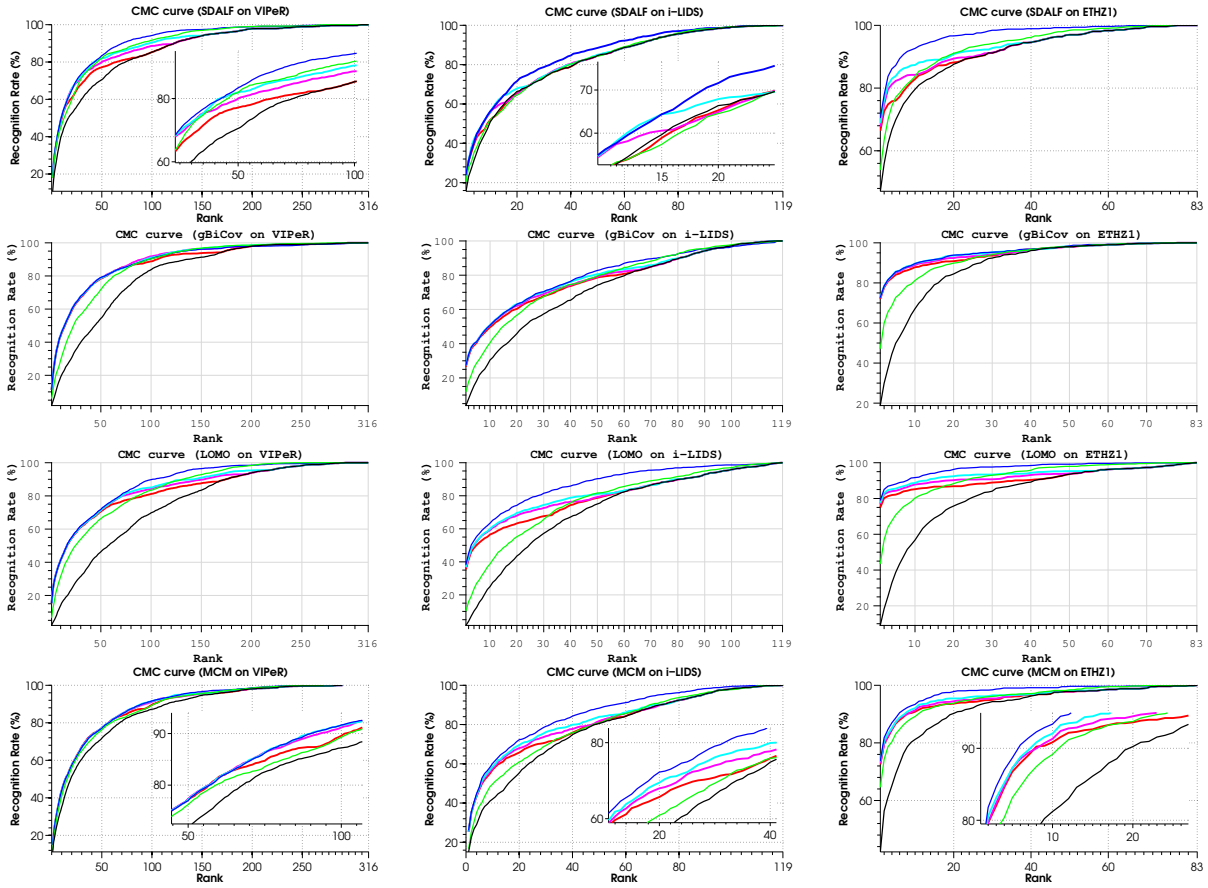

Fig. 5: CMC curves of three-stage systems. Black: first stage; green: second stage; blue: third stage (original descriptor); red, pink, and cyan: three-stage systems with $\beta = 0.3, 0.4, 0.5$, respectively. Enlarged version of plots with very close CMC curves are shown for better visualization.

pp. 63–73

14 Pudil, P., Novovicova, J., Blaha, S., Kittler, J. 'Multistage pattern recognition with reject option'. In: Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on. (IEEE, 1992. pp. 92–95

15 Kaynak, C., Alpaydin, E. 'Multistage cascading of multiple classifiers: One man's noise is another man's data'. In: ICML. (Citeseer, 2000. pp. 455–462

16 Trapeznikov, K., Saligrama, V., Castañón, D.: 'Multi-stage classifier design', *Machine learning*, 2013, **92**, (2-3), pp. 479–502

17 Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H. 'Person re-identification by descriptive and discriminative classification'. In: Image Analysis. (Springer, 2011. pp. 91–102

18 Metternich, M.J., Worring, M.: 'Track based relevance feedback for tracing persons in surveillance videos', *Computer Vision and Image Understanding*, 2013, **117**, (3), pp. 229–237

19 Liu, C., Loy, C.C., Gong, S., Wang, G. 'Pop: Person re-identification post-rank optimisation'. In: Computer Vision (ICCV), 2013 IEEE International Conference on. (IEEE, 2013. pp. 441–448

20 Wang, Z., Hu, R., Liang, C., Leng, Q., Sun, K. 'Region-based interactive ranking optimization for person re-identification'. In: Advances in Multimedia Information Processing–PCM 2014. (Springer, 2014. pp. 1–10

21 Huang, S., Gu, Y., Yang, J., Shi, P. 'Reranking of person re-identification by manifold-based approach'. In: Image Processing (ICIP), 2015 IEEE International Conference on. (IEEE, 2015. pp. 4253–4257