POP: Person Re-Identification Post-Rank Optimisation

Chunxiao Liu[1], Chen Change Loy[2], Shaogang Gong[3], Guijin Wang[1]
[1] Dept. of Electronic Engineering, Tsinghua University, China
[2] Dept. of Information Engineering, The Chinese University of Hong Kong
[3] School of EECS, Queen Mary University of London, UK
lcx08@mails.tsinghua.edu.cn, ccloy@ie.cuhk.edu.hk
sgg@eecs.qmul.ac.uk, wangguijin@tsinghua.edu.cn

Abstract

Owing to visual ambiguities and disparities, person re-identification methods inevitably produce suboptimal rank-list, which still requires exhaustive human eyeballing to identify the correct target from hundreds of different likely-candidates. Existing re-identification studies focus on improving the ranking performance, but rarely look into the critical problem of optimising the time-consuming and error-prone post-rank visual search at the user end. In this study, we present a novel one-shot Post-rank OPtimisation (POP) *method, which allows a user to quickly refine their search by either "one-shot" or a couple of sparse negative selections during a re-identification process. We conduct systematic behavioural studies to understand user's searching behaviour and show that the proposed method allows correct re-identification to converge 2.6 times faster than the conventional exhaustive search. Importantly, through extensive evaluations we demonstrate that the method is capable of achieving significant improvement over the state-of-the-art distance metric learning based ranking models, even with just "one shot" feedback optimisation, by as much as over 30% performance improvement for rank 1 re-identification on the VIPeR and i-LIDS datasets.*

1. Introduction

For person re-identification (re-id), a probe image serves as a query to be compared against a gallery that consists of images of different individuals captured at distributed locations at different time. Typically, a rank list of possibly hundreds of matched likely-images are returned by an appearance-based matching method. The final judgement is left to the end user, who needs to inspect the list and manually search for the correct match to the query. Existing re-identification methods generally assume the rank list being good enough for decision making. In reality, such a ranking list is far from good and necessarily suboptimal.

Figure 1. Human-in-the-loop re-identification is needed to resolve the inherent visual ambiguities and disparities caused by different camera view orientations, occlusion, and lighting variations.

We wish to address this problem of *person re-id post-rank optimisation*, as it is both non-trivial and critical for making the existing re-identification pipeline viable for any real-world practical applications. There are two reasons for such considerations:

Visual ambiguities and disparities - In the context of person re-identification, the visual samples are ambiguous, *i.e*. the same person can look very different and different people can look very alike under different camera views, lighting variations, and occlusion (Fig. 1). Within the vast amount of likely candidates, there may be only one correct target. This problem is perhaps uniquely so for re-identification, whilst less so for general object search in the context of image indexing and search, of which the retrieved images have strong inter-category visual differences and intra-category similarities, are well segmented and largely exhibited from similar views.

Off-line learning scalability - The performance of current distance learning based ranking approaches to person re-identification remain low [19, 26, 13, 17, 16, 25], *e.g*. $\leq$30% recognition rate at rank 1 on the popular VIPeR dataset even with person probe images manually and carefully cropped. A key factor that contributes to the poor re-

This research was partially supported by NSFC 61271390 and Vision Semantics Ltd.

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.62

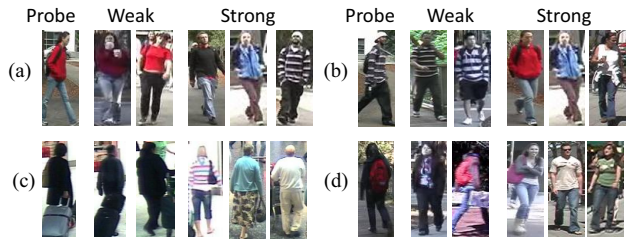

Figure 2. Examples of user negative selections.

sults is the lack of sufficient labelled pairs of training samples to cover diverse appearance variations from unknown changes in viewing conditions, leading to suboptimal learning of the ranking function between camera views. In addition, there is currently no effective mechanism to utilise additional information to further improve model ranked re-id results. Owing to these difficulties, a rank list is inevitably suboptimal. Our experiments show that in each post-rank search, a user spends an average of 45 seconds to identify a true match given a machine generated rank-list (i.e. post-rank) for VIPeR dataset (316 gallery images) (Sec. 5). It is unrealistic to expect an operator to scroll down hundreds of images to find a possible true re-identification in a practical system.

The main contribution of this study is that we formulate a systematic framework for re-id post-rank optimisation, largely unaddressed by the existing person re-identification literature. We introduce a new *one-shot Post-rank OPtimisation (POP)* model for very fast post-rank re-identification convergence with significant increase in re-id accuracy. Specifically, our method aims to minimise human-in-the-loop effort by *one-shot* negative feedback selection. That is, a user only needs to select *a single* strong negative feedback, and optionally a few weak negatives, to trigger an automated refinement of the suboptimal rank list. A strong negative is a highly ranked, but confusing match in a machine generated suboptimal rank list with clear visual dissimilarity to the probe image, whilst a weak negative is a visually similar but wrong match in the same rank list (Fig. 2). We formulate a new visual expansion model that not only synthesises pseudo-samples to complement the sparse negative selection, but also compute a generic mapping of visual change between different camera views. In addition, we introduce an incremental affinity graph construction for propagating sparse belief accumulated from human-in-the-loop negative mining. In essence, the proposed model combines sparse human negative feedback on-the-fly to steer automatic selection of more relevant re-identification features.

We show in Sec. 6 that our model not only improves 2.6 times of search efficiency compared to the typical exhaustive search strategy, but also brings about as much as over 30% performance improvement for rank 1 re-identification over current distance metric learning and ranking models. This is based on "one shot" user negative selection only, and evaluated extensively using both the VIPeR and i-LIDS benchmark datasets.

2. Related Work

Post-rank optimisation for re-id is relatively unexplored in the person re-identification literature. One related study in [12] attempted to refine the rank list but their study does not model the process of enabling human-in-the-loop for optimising the suboptimal rank list with only sparse feedback, down to one-shot. Ali *et al*. [1] proposed to exploit human supervision during a visual search process. In contrast to the proposed one-shot POP model in this study, their interactive scheme requires a user to provide both multiple similar and dissimilar examples, which is not always practical or accessible. Another related work [18] requires explicit relative feedback in image classifier training to diffuse the label to unlabelled images. Their interactive scheme demands more detailed and specific feedback, which may not be feasible in the context of person re-identification when visual cues are often of low-resolution, ambiguous, and lack of relative details.

In a wider context, studies in [2, 7] primarily address a different problem, i.e. face recognition in multimedia domain with feedback for query expansion in continuously tracked faces, a significantly more constrained problem when compared to person re-identification by a single image (see Fig. 1). Continuously tracked facial images mostly undergo smooth appearance changes under strong space-time closed-world constraints, with minimal or no occlusion and very rich data for model learning. In contrast, the person re-id challenge is concerned with a single pair of image association in a totally unconstrained open-world environment.

Our negative mining concept is related to human relevance feedback mining in generic image search and retrieval. However, methods designed for generic inter-class image categerisation are not directly applicable to the person re-identification problem. This is not only due to the visual ambiguity challenge unique to person re-id scenarios as discussed in Sec. 1 and illustrated in Fig. 1, but also because some key underlying assumptions required by most generic image search and retrieval techniques are no longer applicable in the case of person re-id. They are: (1) top-ranked positive images are visually consistent to the probe (no visual ambiguities) [24, 10], (2) those positive images often form the largest cluster [28], or (3) sufficient positive samples can be gathered through text keyword expansion [21]. Returning only probe-relevant images at the top rank cannot be guaranteed in person re-id due to visual variations across camera views. A true positive person re-id match does not necessarily forms a large cluster in the gallery set, in the contrary it is often sparse. Keyword expansion is not applicable to person re-identification scenarios.

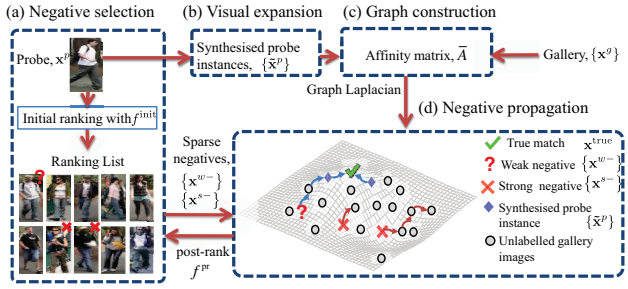

Figure 3. An overview of the proposed one-shot Post-rank OPtimisation (POP) model for person re-identification.

3. Optimising Post-Rank by Negative Mining

3.1. Human-in-the-Loop Negative Mining

Let us consider solving the following person re-identification problem. Given a probe image to be matched against an *unlabelled* gallery set, a ranking function generates a suboptimal rank list of the gallery set according to each gallery image's likelihood to be a true match of the probe image. There may exist only one true match and there is no guarantee that the ranking function is able to place it in the top ranks. All other samples in the gallery space are considered as negatives, which can be divided into two negative types (Fig. 2): (1) Strong negatives - highly ranked gallery images that are visually clearly dissimilar to the probe image. Flagging out one of them may help explaining away many other false matches. (2) Weak negatives - albeit not the true match, these highly ranked negative gallery images are visually similar to the probe image. They could be good candidates for disambiguating visual uncertainties and optimising the initial ranking function. We wish to formulate a model to best exploit human-in-the-loop feedback for post-rank optimisation.

More precisely, for each image we extract a d-dimensional feature vector, denoted by $\mathbf{x} = (x_1, \ldots, x_d)^{\mathsf{T}} \in \mathbb{R}^d$. Given a probe instance, $\mathbf{x}^p$, we assume an initial ranking function f^{init} is available (e.g. [19, 26]) to compute an initial score vector $\mathbf{s}^{\text{init}} = \left(s_1^{\text{init}}, \ldots, s_n^{\text{init}}\right)$ to rank the n unlabelled gallery images $\{\mathbf{x}_i^g\}_{i=1}^n$. If the initial ranking function fails to return the true match $\mathbf{x}^{\text{true}}$ in the top N ranked candidates, we wish to learn a post-rank function f^{pr} for rank re-ordering. This problem is solved by the following procedure, with an overview in Fig. 3:

(a) A user selects one (any) strong negative from the top N ranked instances, denoted as $\mathbf{x}^{s-}$ [1].

(b) For learning the post-rank function, we also require positive sample(s) in addition to the user selected negative sample. To that end, visual expansion is computed to synthesise one or more instances of the probe image ($\bar{\mathbf{x}}^p$) in the gallery view (Sec.3.2).

(c) An affinity graph weighted by an affinity matrix $\bar{A}$ is constructed to capture the appearance similarities among all the images in the gallery view, including both the original gallery instances and the synthesised probe instances (Sec.3.3).

(d) This sparse negative information obtained from the user is propagated to their nearby neighbours in the gallery view via the above weighted affinity graph (Sec.3.4). Through this process, the post-rank function f^{pr} is learned. The initial score $\mathbf{s}^{\text{init}}$ is combined by weighted sum with those obtained from f^{pr} to produce a new set of scores. All instances in the initial rank list are then re-ordered based on the new scores.

3.2. Cross-Camera View Visual Expansion

Learning a post-rank function for rank re-ordering requires both labelled negative and positive data. Clearly, a single strong negative selected by user is insufficient for this purpose. A plausible solution is by synthesising some pseudo positive-labelled samples through a process of visual expansion. However, this is computationally non-trivial in the context of person re-identification. As discussed in Sec. 2, existing visual expansion methods are not directly applicable since visual consistency in top-ranked images cannot be guaranteed. Moreover, owing to potentially large feature inconsistency between different camera views, the probe image itself from the probe camera view cannot be readily used as a positive sample in the gallery view.

To resolve this problem, we specifically design a regression forest [4] based visual expansion method. The regression forest is well-suited to our problem due to its robustness in learning non-linear mapping between high-dimensional re-id visual features. Moreover, the nature of it being an ensemble of trees allows efficient random permutation in the predictors to synthesise one or more samples that resemble the probe's appearance as pseudo positive-labelled data in the gallery view.

Specifically, the visual variations between a probe and a gallery camera view are accounted by the multi-output regression forest, with T_r trees, through learning an appearance mapping space

$$M : \mathbf{x}^p \rightarrow \mathbf{x}^g \in \mathbb{R}^d, \tag{1}$$

from a set of paired training instances extracted from cross-camera views (Fig. 3(b)). A synthesised probe instance can then be generated as follows

[1] Although the user also has the option to select more than one strong negatives $\{\mathbf{x}^{s-}\}$ and a couple of weak negatives $\{\mathbf{x}^{w-}\}$, we will show in Sections 5 and 6 that a single strong negative is far more likely to be selected by a human user (visually distinct and intuitive) than weak negatives (visually subtle) in an on-the-fly feedback process. We also show in comparative experiments in Section 6 that any performance advantage gained from additional multiple negative feedback over a single one-shot negative feedback is insignificant as a result of post-rank optimisation.

$$\tilde{\mathbf{x}}^p = \sum_{t=1}^{T_s} M_{\pi_t}(\mathbf{x}^p), \quad (2)$$

where $T_s = \frac{2}{3}T_r$[2], and M_t is the regression predictor for the t-th regression tree. The subscript $\{\pi_1, \ldots, \pi_{T_s}\}$ is a randomly sampled index and $\pi = \{1, \ldots, T_r\}$. This process can be repeated to generate more synthesised probe instances if desired.

3.3. Incremental Construction of Affinity Graph

We use $\{\mathbf{x}^{s-}\}$ to represent dissimilarity and $\{\tilde{\mathbf{x}}^p\}$ to indicate similarity. If $\{\mathbf{x}^{w-}\}$ are selected, they can be treated as positives due to the fact that they are visually similar to the probe[3]. To that end, we shall describe how to propagate the sparse labelled samples to the large quantity of unlabelled gallery set so to avoid the need for labelling exhaustively the gallery set. This process of transduction via an affinity graph is facilitated by first constructing an affinity graph of the unlabelled gallery set.

In contrast to existing graph-based methods [23, 14, 11, 16], we exploit clustering forest [4, 5, 27] to discover the distances between images in order to address the inherent noise in the re-id problem. The use of a clustering forest is advantageous to solving this re-id problem in two aspects: (1) its implicit feature selection mechanism is beneficial to mitigating noisy visual features, and (2) it offers scalable and tractable solution to our incremental graph construction requirement so to accommodate varying number of selected negatives accumulating on-the-fly. Note that the unsupervised clustering forest differs from the supervised regression forest we used in Sec. 3.2 for visual expansion.

Let us first describe how to construct a graph for the gallery instances $\{\mathbf{x}^g\}$ excluding the synthesised probe instances $\{\tilde{\mathbf{x}}^p\}$, which are not part of the gallery and also not a constant number depending on user negative selection choices. We shall return to the case of including synthesised positives later. Our clustering forest is an ensemble of T_c trees (Fig. 3(c)), each of which defines a partition of the inputs $\mathbf{x}^g$ at its leaves, $q(\mathbf{x}^g) : \mathbb{R}^d \rightarrow \mathcal{Q} \subset \mathbb{N}$, where q represents a leaf index and $\mathcal{Q}$ is the set of all leaves in a given tree. In the t-th tree, the distance of $\mathbf{x}_i^g$ and $\mathbf{x}_j^g$ is

$$\mathrm{dist}^t\left(\mathbf{x}_i^g, \mathbf{x}_j^g\right) = \begin{cases} 0 & \text{if } q(\mathbf{x}_i^g) = q(\mathbf{x}_j^g) \\ \infty & \text{otherwise} \end{cases}. \quad (3)$$

We then collect the pairwise distances of all gallery instances to construct an affinity matrix $A^t \in \mathbb{R}^{n \times n}$ of that tree, with each element A_{ij}^t given as

$$A_{ij}^t = \exp^{-\mathrm{dist}^t\left(\mathbf{x}_i^g, \mathbf{x}_j^g\right)}. \quad (4)$$

Intuitively, we assign affinity=1 (distance=0) to samples $\mathbf{x}_i^g$ and $\mathbf{x}_j^g$ if they fall into the same leaf node, and affinity=0 (distance=∞) otherwise. To obtain a smooth forest affinity matrix, we compute the final affinity matrix as $\bar{A} = \frac{1}{T_c}\sum_{t=1}^{T_c} A^t$. The affinity is then used to weigh the edges in an k-NN graph.

Let us now consider the case for including synthesised positives in the construction of the affinity graph. Recall that our method is designed to need only a single strong negative to re-order the rank. Nevertheless, a user has the option to select more negatives in more than one round of feedback, if necessary and desired. To maintain a balance in positive-negative data for the post-rank function learning, the model needs to generate equal number of synthesised positive probe instances $\{\tilde{\mathbf{x}}^p\}$ as pseudo positive-labelled data in the gallery view. Thus, the number of $\tilde{\mathbf{x}}^p$ can vary depending on the number of negatives selected by a user cumulatively. Constructing a new graph from scratch catering for each increase in the number of $\tilde{\mathbf{x}}^p$ is infeasible, since it involves a complexity order of $\mathrm{O}\left((n+\tilde{n})^2\right)$, where $\tilde{n}$ is the number of $\tilde{\mathbf{x}}^p$. A more tractable approach is to first build a graph using the gallery data alone without the additional synthesised positives, and then expand it to accommodate the additional synthesised probe instances, as follows.

First, we compute the affinity between $\{\tilde{\mathbf{x}}^p\}$ and all the existing gallery instances $\{\mathbf{x}^g\}$. Benefited from the tree structure of clustering forest, the affinity computation is efficient. In particular, since the index of each gallery instances is stored in the leaf nodes during the forest construction, we can compute $\mathrm{dist}^t\left(\mathbf{x}_i^g, \tilde{\mathbf{x}}_j^p\right)$ by checking on which leaf node an $\tilde{\mathbf{x}}_j^p$ fall in a tree. For distances between synthesised probe instances, we compute $\mathrm{dist}^t\left(\tilde{\mathbf{x}}_i^p, \tilde{\mathbf{x}}_j^p\right) = \min\left\{\mathrm{dist}^t\left(\tilde{\mathbf{x}}_i^p, \mathbf{x}_j^g\right) | j = 1, \ldots, n\right\}$. Second, with the new set of distances we can then expand the old affinity matrix from $\bar{A} \in \mathbb{R}^{n \times n}$ to $\bar{A} \in \mathbb{R}^{(n+\tilde{n}) \times (n+\tilde{n})}$, followed by affinity normalisation. New nodes corresponding to $\{\tilde{\mathbf{x}}^p\}$ are subsequently added to the original k-NN graph.

3.4. Sparse Negative Propagation over Graph

After constructing the affinity graph, we diffuse the sparse negative and synthesised positive information over the graph to all other gallery instances. First, we order the selected negatives and synthesised probe instances into the first l labelled samples $\mathcal{L}$, followed by the remaining u gallery instances as unlabelled samples $\mathcal{U}$, *i.e.*

$$\mathcal{L} = \left\{\mathbf{x}^{s-}\right\} \cup \left\{\mathbf{x}^{w-}\right\} \cup \left\{\tilde{\mathbf{x}}^p\right\} \quad (5)$$

$$\mathcal{U} = \left\{\mathbf{x}^g\right\} \setminus \left(\left\{\mathbf{x}^{s-}\right\} \cup \left\{\mathbf{x}^{w-}\right\}\right)$$

$$y = \begin{cases} +1 & \text{if } \mathbf{x} \in \left\{\mathbf{x}^{w-}\right\} \cup \left\{\tilde{\mathbf{x}}^p\right\} \\ -1 & \text{if } \mathbf{x} \in \left\{\mathbf{x}^{s-}\right\} \end{cases}. \quad (6)$$

Here we accommodate the possibility of a user wanting to select some weak negatives. Otherwise $\{\mathbf{x}^{w-}\} = \emptyset$.

Second, to propagate negative information from $\mathcal{L}$ to $\mathcal{U}$, we consider the following optimisation problem similar to

[2]This fraction is typical in random forest bootstrap training [4].

[3]Using similar examples (here the weak negative) as positive is also explored in label sharing [6] and example sharing [20].

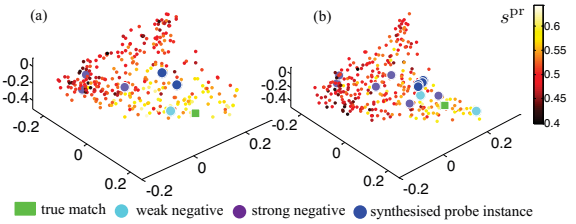

Figure 4. Effects of negative accumulation: (a) three-dimensional embedding of gallery images obtained using multi-dimensional scaling after the first round of negative selection, (b) the embedding after the second round. The gallery images are colour coded according to their new ranking score. The shrinking region of bright yellow colour indicates the effectiveness of negative mining in demoting initial false matches.

Laplacian SVM [3]:

$$f^{\text{pr}} = \underset{f\in\mathcal{H}_K}{\text{argmin}} \frac{1}{l}\sum_{i=1}^{l} \max(1-y_i f(\mathbf{x}_i),0) + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2, \quad (7)$$

where $f^{\text{pr}} : \mathbf{x} \rightarrow \mathbb{R}$ is the post-rank function. The first term defines a hinge loss on the sparse labelled data. The parameter λ_A enforces a smoothness condition on the solution, and $\|f\|_K^2$ denotes the norm of the function in Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ [3]. The parameter λ_I controls the intrinsic regulariser $\|f\|_I^2$, which enforces the similar/dissimilar labels of nearby gallery instances with respect to the affinity graph to be close. Specifically $\|f\|_I^2 = \mathbf{f}^{\mathsf{T}} L \mathbf{f} = \sum_{i=1}^{l+u}\sum_{j=1}^{l+u} \bar{A}_{ij}(f(\mathbf{x}_i)-f(\mathbf{x}_j))^2$ where $L = D - \bar{A}$, and D represents a diagonal matrix with $D_{ii} = \sum_j \bar{A}_{ij}$. In this study, we use $L = D^{-1/2}\bar{A}D^{-1/2}$ instead since it provides certain theoretical guarantees and perform well in many tasks [22].

Third, we solve Eqn. (7) to derive the Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{l+u})^{\mathsf{T}}$ and the bias term b, by using the Newton's method [3]. Finally, the estimated relevance of an unlabelled gallery instance $\mathbf{x}_j^g$ to the probe is computed as

$$s_j^{\text{pr}} = f^{\text{pr}}(\mathbf{x}_j^g) = \sum_{i=1}^{l+u} \alpha_i K(\mathbf{x}_i, \mathbf{x}_j^g) + b, \quad (8)$$

where kernel $K(\cdot,\cdot)$ denotes a radial basis function in our implementation. We yield the final matching score as

$$\mathbf{s} = (1-\beta)\mathbf{s}^{\text{init}} + \beta \mathbf{s}^{\text{pr}}, \quad (9)$$

where $\mathbf{s}^{\text{pr}} = (s_1^{\text{pr}}, \ldots, s_n^{\text{pr}})$. The parameter β balances the influence between initial ranking and user feedback selections.

3.5. Negative Accumulation

After each round of negative mining, we add new negative selections to a cumulated strong negative sets collected from previous rounds (or also weak negative sets if weak negatives were selected). Figure 4 shows an example for the effect of feedback accumulation in two rounds of negative mining. As more negatives are accumulated, the classification boundary is refined, increasing the separation between the true match and other strong negatives. The above negative accumulation are repeated together with the negative mining steps (Sec. 3.1) until the true match is found in the top ranks, or terminates after a pre-defined number of rounds.

4. Experimental Settings

Datasets - Two widely employed benchmarking datasets VIPeR [9] and i-LIDS [26] were used for evaluation. The VIPeR dataset contains 632 persons, each of which has two images captured in outdoor views. The dataset is challenging due to drastic appearance difference between most of the matched image pairs caused by viewpoint variations and large illumination changes at outdoor environment. The i-LIDS dataset was captured in a busy airport arrival hall using multiple cameras. It contains 119 people with a total of 476 images, with an average of four images per person. Apart from the illumination changes and pose variations, many images in this dataset are also subject to severe inter-object occlusions.

Features - Similar to [19, 26, 15, 16], we partitioned an image equally into six horizontal stripes, and extracted a mixture of colour (RGB, HSV and YCbCr) and texture histograms (8 Gabor filters and 13 Schmid filters), forming a 2784-dimensional feature vector for each image.

Implementation details - We set T_c = T_r = 200 for the forest size. The depth of the forest was automatically discovered by specifying the minimum forest node sizes, *i.e.* 1 for the clustering forest and 5 for the regression forest. The number of nearest neighbours in a k-NN graph was chosen as 20. We set $\lambda_A = 0.1$, $\lambda_I = 0.1$, variance in kernel $K(\cdot,\cdot)$ to 1.5, score fusion parameter $\beta = 0.8$ through cross-validation and kept them fixed in all the experiments. Good performance is consistently observed when we set β in the range of (0.8,1).

Evaluation settings - The matching performance was measured using the averaged cumulative match characteristic (CMC) curve [9] over 10 trials. We selected all the images of p person to build the test set. The remaining data was used for training an initial ranking function and the regression forest. The value p was set to 316 for VIPeR and 50 for i-LIDS. In the test set of each trial, we randomly chose one image from each person to set up the test gallery set and the remaining images were used as probe images. Note that for the i-LIDS dataset, 50 images in the gallery set were insufficient to construct the intrinsic regulariser $\|f\|_I^2$ in Eqn. (7). Thus, we randomly selected 300 images from VIPeR to help in computing the i-LIDS's intrinsic regulariser. During the empirical evaluation, these 'borrowed' examples were never presented to the participants.

5. Behavioural Studies on Post-Rank Search

There are two user studies with the purpose of: (1) understanding the tendency of a user in selecting either strong or weak negative, and (2) quantifying and comparing the search efficiency of the conventional exhaustive search and POP in the hardest case, when user selects only a single strong negative for post-rank optimisation (one-shot).

A total of 10 volunteers were invited for the first study. They were asked to manually annotate the weak and strong negatives ranked by an off-line ranking model given a set of random probe images. It is evident from Table 1 that the proportion of weak and strong negatives are extremely imbalanced with the strong negatives outnumbers the weak negatives significantly. We found that different factors may affect human judgement in the negative selection process, *e.g.* colours (Fig. 2(a)), texture (Fig. 2(b)), accessories such as a luggage case (Fig. 2(c)), or even some ambiguous visual traces (Fig. 2(d)). Overall, these results suggest that the relatively more salient strong negatives are more likely to be selected by a user during a post-rank feedback selection process. This raises the question on how the POP model performs given a single strong negative feedback (i.e. one-shot) as compared to its performance given multiple weak negatives as feedback. We shall evaluate this in Sec. 6.

In the second search efficiency study, a total of 15 participants were invited, each of whom was assigned 10 probes from the VIPeR dataset. The users were shown the initial matching results by $\ell 1$-norm, and were asked to perform one-shot strong negative selection from the top 15 ranked results. They were allocated a maximum of 3 rank feedback rounds with one strong negative selection each. If the true match cannot be promoted into the top 15 ranks by the model after the maximal 3 rounds of one-shot post-rank optimisation, the users were asked to continue with an exhaustive visual search to find the true match. Their search time is automatically recorded. Similar experiment was conducted on using exhaustive search for comparison.

Figure 5 depicts several examples of actual user interactions during the post-rank optimisation process. One can observe that the POP is effective in demoting candidates who have similar appearances to the selected strong positives. For instance, as shown in Fig. 5(b), when a user selected the first candidate as strong negative, both the first and second candidates who were wearing brown jackets were removed from the top ranks. Fig. 5(c) shows a failure case where selecting one strong negative is insufficient to resolve the visual ambiguity, since the true match experiences large appearance variation due to viewpoint change. Figure 6 shows the search time versus the initial rank. The search time of POP is reduced by 2.6 times on average as compared to conventional exhaustive search by ranking. That is from 50.24±24.55 seconds to 19.44±14.51 seconds. These results suggest that the proposed POP model is able to significantly improve the search efficiency.

dataset	n(gallery)	weak	strong	unlabelled
VIPeR	632	1.73%	78.10%	20.17 %
i-LIDS	119	1.04%	62.79%	36.17 %

Table 1. Proportion of user selected strong and weak negatives.

Figure 5. Examples of user feedback on-the-fly. The probe and the true match are highlighted respectively with red and green bounding boxes. In the middle we show the returned top 15 ranked results. The selected strong negative is denoted by a red cross.

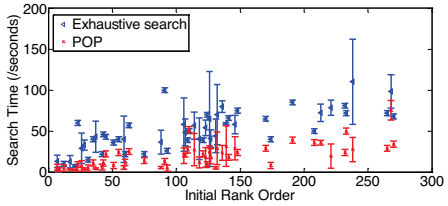

Figure 6. Search time (in second) comparison, between exhaustive search and POP search, both of which are initialised by $\ell 1$-norm matching (best viewed in colour).

6. Comparative Evaluations

POP vs. $\ell 1$-norm, RankSVM, PRDC, MCC – First we evaluate the benefits of POP on existing ranking based person re-identification methods using $\ell 1$-norm [26], RankSVM [19], PRDC [26] and MCC [8], which are among the top performers in re-id. In each round, the negative selection was performed on the first N ranked images, $N = 15$ for the VIPeR dataset and $N = 10$ for the i-LIDS dataset due to its relatively smaller size. We treat the negative selections collected offline from the first behaviour study (Sec.5) as ground truth feedbacks from users. This is to automate the experiments for systematic evaluation of our approach with cross validation. Despite the negative selection was performed without a live user in the loop, the experiments were still using the real feedback from users. This testing

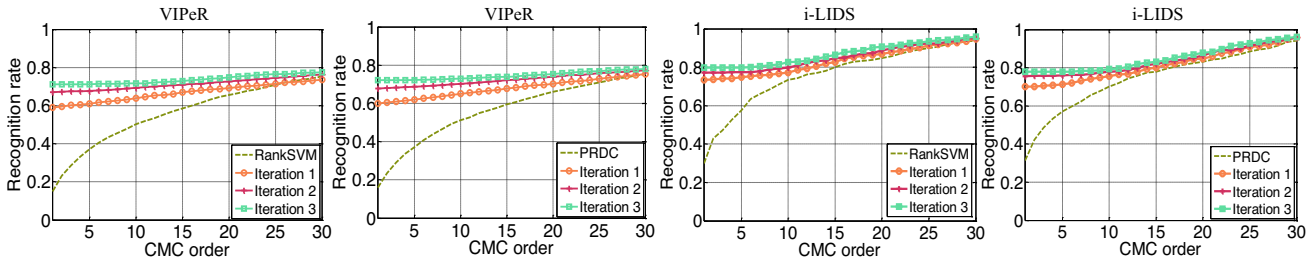

Figure 7. POP post-rank optimisation vs. two re-identification models RankSVM and PRDC (with three rounds of feedback).

Initial Ranking	VIPeR								i-LIDS							
	one-shot				multi-shots				one-shot				multi-shots			
	0	Round 1	R2	R3	0	R1	R2	R3	0	R1	R2	R3	0	R1	R2	R3
ℓ1-norm	9.43	31.90	42.88	47.56	9.43	30.41	44.21	50.13	29.60	67.60	73.20	75.60	29.60	67.60	75.40	81.80
RankSVM [19]	14.87	59.05	67.06	71.08	14.87	58.48	67.85	71.58	29.80	73.40	77.20	79.80	29.80	73.40	79.40	82.40
PRDC [26]	16.01	59.91	67.88	72.03	16.01	59.49	68.35	72.22	31.40	70.20	75.60	78.00	31.40	70.20	77.00	80.00
MCC [8]	17.85	60.13	64.08	66.87	17.85	60.06	63.64	66.20	30.00	69.80	73.60	76.60	30.00	69.20	74.80	80.40

Table 2. Rank-1 recognition rate(%) vs. the number of feedback round on VIPeR and i-LIDS.

protocol was applied for all the experiments reported below.

We conducted both 'one-shot' and 'multi-shots' experiments. The one-shot experiment depicted an extremely sparse feedback scenario, where only one strong negative within the top N ranked images was selected in a round. In a multi-shots scenario, the model was presented with multiple labelled negatives (strong and weak) by a user. The maximum number of strong negatives was set to 5 assuming that the users do not bother to annotate more. Figure 7 and Table 2 show that the recognition rate is remarkably improved with just 1 round of one-shot feedback. Specifically, the rank-1 average recognition rates are boosted by 38.22% and 40.05% on VIPeR and i-LIDS respectively for all four different initial ranking models (ℓ1-norm, RankSVM, PRDC and MCC). With feedback increased to three rounds, the performance improves monotonically and converges. It is worth pointing out that even though RankSVM, PRDC and MCC already achieve a good initial recognition as compared to ℓ1-norm, notable performance gains are achieved after post-rank optimisation by POP.

The performance comparisons between one-shot and multi-shot negative selections are reported in Table. 2. The one-shot negative selection in just one feedback round yields stable and competitive results with no obvious degradation in comparison to the multi-shot multi-rounds feedback, indicating the effectiveness of one-shot post-rank optimisation.

POP vs. other Post-Rank Models – We also compared POP against other post-rank models for generic image search and retrieval tasks, including:

1. NPRF [24] : An SVM is trained by using top-ranked images as positive examples and bottom-ranked images as negative examples.
2. PRF [10]: An one-class SVM is trained by using the top-ranked images as positive examples.
3. EMR [23]: A graph-based ranking method. It uses Euclidean distance to construct the affinity matrix and optimises a ranking function with least square regression. We treat it as a post-rank method by feeding it with the same weak/strong negative selections as POP.

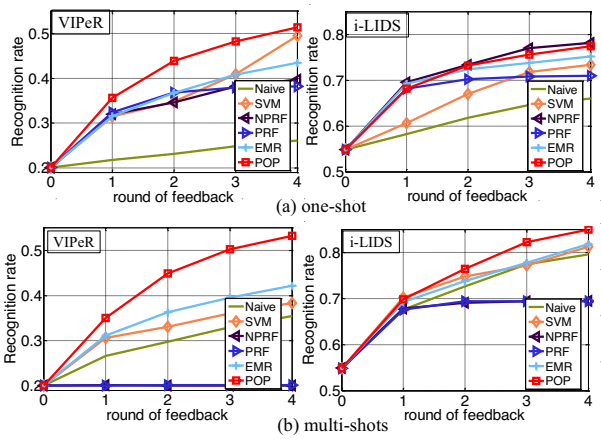

Figure 8. Post-rank optimisation by POP vs. Naïve feedback, and other image retrieval models including NPRF [24], PRF [10], and EMR [23], with ℓ1-norm as the initial ranking function. The y-axis shows the recognition rate at rank-5 along with the increment of feedback round. (a) one-shot, (b) multi-shots.

In addition, we implemented two baseline approaches: (1) a naïve feedback method which simply demotes the strong negatives to the bottom of the ranking list in each round; (2) a SVM approach using the strong negatives and synthesized positive examples for training. For NPRF and PRF, we applied their default strategy for selecting positive and negative samples, and RBF as their SVM kernel, with parameters determined by cross-validation. For EMR, we used the default settings from the authors' code[4]. The ℓ1-norm distance measure was chosen as initial ranking function. Figure 8 shows the comparative rank-5 recognition

[4]http://eagle.zju.edu.cn/~binxu/publication/EMR/EMR.htm

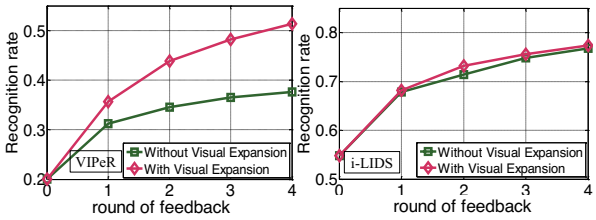

Figure 9. Benefits from visual expansion. $\ell 1$-norm is the initial ranking function. The y-axis shows the recognition rate at Rank-5 along with the increment of feedback round.

rates from all the models, on both one-shot and multi-shots evaluations. NPRF, PRF and the naïve feedback are generally poor in boosting the recognition rate on VIPeR dataset, suggesting that the use of top-ranked images as *positive* feedback samples can lead to erroneous post-rank results in a re-identification task. Both POP and EMR are able to achieve notable gain and are better than SVM method, indicating the strength of label propagation in a graph. POP outperforms EMR by 7.94% (one-shot) and 11.01% (multi-shots) for Rank-5 results on the VIPeR dataset after 4 rounds of post-rank human-in-the-loop process. POP also outperformed EMR by 4.00% (multi-shots) for Rank-5 on the i-LIDS dataset after 4 rounds of feedback, whilst the two giving comparable results for one-shot feedback. The better performance of POP over EMR suggests the more effective propagation of negatives over the clustering-forest based affinity graph, rather than the Euclidean-based graph.

Benefits from Visual Expansion – We further evaluated the additional benefits from visual expansion, with $\ell 1$-norm for initial ranking. We focused on the one-shot case. To prepare the baseline without visual expansion, we randomly selected one weak negative image from the top N ranks ($N = 15$ for VIPeR, 10 for i-LIDS) to pair with the one-shot strong negative. Figure 9 shows that visual expansion improves the recognition rate of POP from 37.66% to 51.39% after 4 feedback rounds on the VIPeR dataset. However, no notable improvement was observed on the i-LIDS dataset. A plausible reason is that the i-LIDS dataset is not partitioned into different camera sets, so learning the mapping space is not as meaningful as in the VIPeR case.

7. Conclusion

We have formulated a systematic framework for re-identification post-rank optimisation, which has been mostly neglected by contemporary person re-identification studies. Systematic behaviour studies and extensive evaluations demonstrated that the proposed POP model not only can improve 3 times of search efficiency over exhaustive search strategy, but also achieves significant improvement over state-of-the-art ranking-based re-id methods, even with just one shot negative selection.

References

[1] S. Ali, O. Javed, N. Haering, and T. Kanade. Interactive retrieval of targets for wide area surveillance. In *ACM MM*, 2010. 2

[2] M. Bäuml, M. Fischer, K. Bernardin, H. K. Ekenel, and R. Stiefelhagen. Interactive person-retrieval in tv series and distributed surveillance video. In *ACM MM*, 2010. 2

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006. 5

[4] L. Breiman. Random forests. *ML*, 45(1):5–32, 2001. 3, 4

[5] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227, 2012. 4

[6] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *ECCV*, 2010. 4

[7] M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Interactive person re-identification in tv series. In *CBMI*, 2010. 2

[8] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *NIPS*, pages 451–458, 2005. 6, 7

[9] D. Gray and H. Tao. Viewpoint invariant pedestrain recognition with an ensemble of localized features. In *ECCV*, 2008. 5

[10] J. He, M. Li, Z. Li, H. Zhang, H. Tong, and C. Zhang. Pseudo relevance feedback based on iterative probabilistic one-class SVMs in web image retrieval. *PCM*, 2005. 2, 7

[11] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. In *ACM MM*, 2004. 4

[12] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011. 2

[13] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 1

[14] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *CVPR*, 2010. 4

[15] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important ? In *ECCV Workshop on Person Re-identification*, 2012. 5

[16] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *ICIP*, 2013. 1, 4, 5

[17] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 1

[18] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*, 2012. 2

[19] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010. 1, 3, 5, 6, 7

[20] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011. 4

[21] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang. Intentsearch: Capturing user intention for one-click internet image search. *TPAMI*, 34(7):1342–1353, 2012. 2

[22] U. Von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008. 5

[23] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, and J. Luo. Efficient manifold ranking for image retrieval. In *SIGIR*, 2011. 4, 7

[24] R. Yan, A. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *ACM MM*, 2003. 2, 7

[25] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 1

[26] W. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *TPAMI*, 35(3):653–668, 2013. 1, 3, 5, 6, 7

[27] X. Zhu, C. C. Loy, and S. Gong. Video synopsis by heterogeneous multi-source correlation. In *ICCV*, 2013. 4

[28] H. Zitouni, S. Sevil, D. Ozkan, and P. Duygulu. Re-ranking of web image search results using a graph algorithm. In *ICPR*, 2008. 2