This CVPR2015 extended abstract is the Open Access version, provided by the Computer Vision Foundation.

Similarity Learning on an Explicit Polynomial Kernel Feature Map for Person Re-Identification

Dapeng Chen[1], Zejian Yuan[1], Gang Hua[2], Nanning Zheng[1], Jingdong Wang[3]
[1] Xi'an Jiaotong University. [2] Stevens Institute of Technology. [3] Microsoft Research

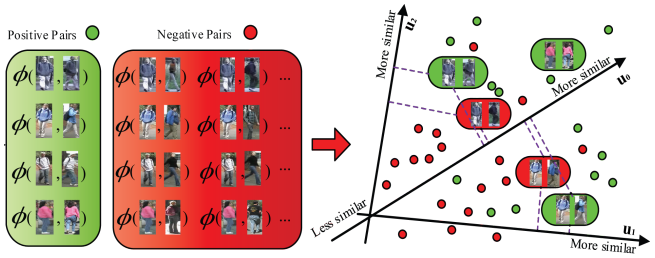

Figure 1: **Flow chart of the proposed method.** We first map the image pairs to the explicit polynomial kernel feature, then train a mixture of similarity functions to discover multiple matching patterns.

1 Introduction

Existing works tackle this problem from two paths. The first one is to design a visual descriptor to handle inter-camera differences in lighting conditions, changes in object orientation and object pose. The second path is to learn a similarity function to suppress inter-camera variations, which our work belongs to.

In this paper, we present an explicit polynomial kernel feature map, which is capable of characterizing the similarity information of all pairs of patches between two images, called soft-patch-matching, instead of greedily keeping only the best matched patch, and thus more robust. Second, we introduce a mixture of linear similarity functions that is able to discover different soft-patch-matching patterns. Last, we introduce a negative semi-definite regularization over a subset of the weights in the similarity function, which is motivated by the connection between explicit polynomial kernel feature map and the Mahalanobis distance, as well as the sparisity constraints over the parameters to avoid over-fitting.

2 Methodology

Similarity Function. We introduce the similarity function $f(\mathbf{x}_1,\mathbf{x}_2)$ for image descriptors $\mathbf{x}_1$ and $\mathbf{x}_2$. The similarity function is performed by an explicit kernel feature map $\phi(\mathbf{z})$ on the concatenated vector $\mathbf{z}=[\mathbf{x}_1^\top \mathbf{x}_2^\top]$. For second-order polynomial kernel $k(\mathbf{z}_1,\mathbf{z}_2)=(\mathbf{z}_1^\top \mathbf{z}_2)^2$, the corresponding feature map $\phi(\mathbf{z})=\phi(\mathbf{z}^\top\mathbf{z})=[\mathrm{vec}(\mathbf{x}_1\mathbf{x}_1^\top)^\top \mathrm{vec}(\mathbf{x}_2\mathbf{x}_1^\top)^\top \mathrm{vec}(\mathbf{x}_1\mathbf{x}_2^\top)^\top \mathrm{vec}(\mathbf{x}_2\mathbf{x}_2^\top)^\top]^\top$. To make the function be symmetric, which is natural for the similarity function, we redefine:

$$\phi(\mathbf{x}_1,\mathbf{x}_2)=[\mathrm{vec}(\mathbf{x}_1\mathbf{x}_1^\top+\mathbf{x}_2\mathbf{x}_2^\top)^\top \ \mathrm{vec}(\mathbf{x}_2\mathbf{x}_1^\top+\mathbf{x}_1\mathbf{x}_2^\top)^\top]^\top. \quad (1)$$

This feature map takes into account the relation between the feature values from the same position and different positions. In the case when feature $\mathbf{x}$ is a patch-wise descriptor of an image (each entry or subvector corresponds to a block of the image), $\mathrm{vec}(\mathbf{x}_1\mathbf{x}_2^\top)$ can be viewed as a concatenation of cross-patch similarities of two images, where the cross-patch similarity is a vector formed by vectorizing the out-product of the patch features. In other words, it matches each patch in one image with all the patches in the other image and all the matching scores are attained as the descriptor, which we call soft-patch-matching, instead of only keeping the best-matched score. This still holds even the descriptor $\mathbf{x}$ undergo certain linear transformation. As we can show that it is equivalent to transforming the matching $\mathrm{vec}(\mathbf{x}_1\mathbf{x}_2^\top)$ into a linear subspace. The similarity function, $f(\mathbf{x}_1,\mathbf{x}_2)$, is usually linear w.r.t. $\phi(\mathbf{x}_1,\mathbf{x}_2)$. To handle different soft-patch-matching patterns, we make non-linear extension using a latent formulation,

$$f(\mathbf{x}_1,\mathbf{x}_2)=\max_{h=1,\dots H} f_h(\mathbf{x}_1,\mathbf{x}_2), \quad (2)$$

where $f_h(\mathbf{x}_1,\mathbf{x}_2;\mathbf{w}_h)=\mathbf{w}_h^\top\phi(\mathbf{x}_1,\mathbf{x}_2)$. Intuitively, the latent formulation aims to discover H representative patterns $\{\mathbf{w}_h\}_{h=1}^H$ and uses the most similar pattern to evaluate the similarity for a pair $(\mathbf{x}_1,\mathbf{x}_2)$ in terms of the inner product.

Regularization. The first regularization is motivated by the connection between explicit polynomial kernel feature map and the Mahalanobis distance. We rearrange $\phi(\mathbf{x}_1,\mathbf{x}_2)=[\phi^1(\mathbf{x}_1,\mathbf{x}_2),\ \phi^2(\mathbf{x}_1,\mathbf{x}_2)]$, where $\phi^1(\mathbf{x}_1,\mathbf{x}_2)=\mathrm{vec}(\mathbf{x}_1\mathbf{x}_1^\top+\mathbf{x}_2\mathbf{x}_2^\top-\mathbf{x}_1\mathbf{x}_2^\top-\mathbf{x}_2\mathbf{x}_1^\top)$ and $\phi^2(\mathbf{x}_1,\mathbf{x}_2)=\mathrm{vec}(\mathbf{x}_1\mathbf{x}_2^\top+\mathbf{x}_2\mathbf{x}_1^\top)$. Accordingly, $\mathbf{w}_h$ is written as $[\mathbf{w}_h^1,\ \mathbf{w}_h^2]$, and the linear function:

$$f_h(\mathbf{x}_1,\mathbf{x}_2)=(\mathbf{w}_h^1)^\top\phi^1(\mathbf{x}_1,\mathbf{x}_2)+(\mathbf{w}_h^2)^\top\phi^2(\mathbf{x}_1,\mathbf{x}_2). \quad (3)$$

We impose the negative semi-definite regularization over $\mathbf{w}_h^1$, $\mathrm{mat}(\mathbf{w}_h^1)\preceq 0$. As it can be considered that we use the negative Mahalanobis distance to measure the similarity. The second regularization is motivated by the assumption that different matching patterns share a common component. We decompose $\mathbf{w}_h=\mathbf{u}_h+\mathbf{u}_0$, and align the weights of the H similarity functions to a common weight vector $\mathbf{u}_0$. The alignment is imposed by a sparsity regularization: $\sum_{h=1}^H\|\mathbf{u}_h\|_1$. In addition, we also impose the sparsity regularization $\|\mathbf{u}_0\|_1$, which is widely used for feature selection.

Problem formulation. The training data for person re-identification can be transformed as follows. Given a set of probe images $\mathcal{X}=\{\mathbf{x}_1,\cdots,\mathbf{x}_N\}$, image $\mathbf{x}_n$ is associated with two sets of gallery images: a positive set $\mathcal{X}_n^+$ composed of the images about the same person with $\mathbf{x}_n$, a negative set $\mathcal{X}_n^-$ composed of the images about different persons. We utilize the triplet loss: $\mathcal{L}(\mathbf{u}_0,\mathbf{u}_1,\dots,\mathbf{u}_H)=\sum_{i=1}^N\sum_{\mathbf{x}_j\in\mathcal{X}_i^+,\mathbf{x}_k\in\mathcal{X}_i^-}[f(\mathbf{x}_i,\mathbf{x}_k)-f(\mathbf{x}_i,\mathbf{x}_j)+1]_+$, where $\mathbf{x}_j\in\mathcal{X}_i^+$ and $\mathbf{x}_k\in\mathcal{X}_i^-$. With the regularization, the objective function for person re-identification is given as:

$$\min_{\mathbf{u}_0,\dots,\mathbf{u}_H}\mathcal{L}(\mathbf{u}_0,\mathbf{u}_1,\dots,\mathbf{u}_H)+\lambda\sum_{h=0}^H\|\mathbf{u}_h\|_1, \qquad s.t.\,\mathrm{M}(\mathbf{u}_h)\preceq 0, h=0,1,\dots,H. \quad (4)$$

where $\mathrm{M}(\mathbf{u}_h)=\mathrm{mat}(\mathbf{u}_h^1)$. $\mathbf{u}_h^1$ is the first half part of $\mathbf{u}_h$. As $\mathbf{w}_h=[\mathbf{w}_h^1,\mathbf{w}_h^2]=[\mathbf{u}_h^1+\mathbf{u}_0^1,\mathbf{u}_h^2+\mathbf{u}_0^2]$, constraints in 4 can derive $\mathrm{mat}(\mathbf{w}_h^1)\preceq 0$.

Optimization. As problem 4 is non-convex, we utilize a EM-like algorithm to iteratively optimize a convex subproblem that is a upper bound of the original problem. Each iteration includes two steps. The first step is to estimate the hidden variables for each positive image pairs, which yeilds the subproblem. The second step is optimize the subproblem by alternating direction method of multipliers (ADMM).

3 Experiment

We evaluate the proposed similarity learning approach for the person re-identification task on three widely-used datasets including VIPER, GRID and CAVIAR4REID, as well as for the face verification task on the LFW dataset. We empirically analyzed how various components in our approach affect the performance, including the influence of explicit polynomial kernel feature map, regularization as well as a mixture of similarity functions.

Methods	Regularization	Feature	H	Loss function	r=1	r=10	r=20
NSS	None	$\phi(\mathbf{x}_1,\mathbf{x}_2)$	6	triplet	24.8	72.9	87.0
SP	Sparse(SP)	$\phi(\mathbf{x}_1,\mathbf{x}_2)$	6	triplet	26.1	76.7	89.5
SD	Semi-definite(SD)	$\phi(\mathbf{x}_1,\mathbf{x}_2)$	6	triplet	**34.4**	**69.5**	**82.7**
F1	SP+SD	$\phi^1(\mathbf{x}_1,\mathbf{x}_2)$	6	triplet	28.8	74.8	86.9
F2	SP+SD	$\phi^2(\mathbf{x}_1,\mathbf{x}_2)$	6	triplet	28.9	77.7	89.2
F3	SP+SD	$\phi^3(\mathbf{x}_1,\mathbf{x}_2)$	6	triplet	13.4	36.6	49.9
Single	SP+SD	$\phi(\mathbf{x}_1,\mathbf{x}_2)$	0	triplet	33.9	81.2	**91.7**
Binary	SP+SD	$\phi(\mathbf{x}_1,\mathbf{x}_2)$	6	binary	33.0	80.6	89.5
Ours	SP+SD	$\phi(\mathbf{x}_1,\mathbf{x}_2)$	6	triplet	**36.8**	**83.7**	**91.7**

Table 1: **The empirical evaluation on the VIPER dataset:** The top-n matching rate of the methods with different configurations about regularization strategy, explicit polynomial kernel feature map, the number of similarity funcion H and the loss function. All the experiments are run 10 times with the same partition. The size of the gallery set is 316.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.