# SALIENCY PREPROCESSING FOR PERSON RE-IDENTIFICATION IMAGES

*Cong Ma*[⋆] *Zhenjiang Miao*[⋆] *Min Li*[⋆]

[⋆] Institute of Information Science, Beijing Jiaotong University, China

## ABSTRACT

In this paper, we propose a preprocessing strategy for refining the pedestrian images in person re-identification(re-id). The accomplishment of the person re-id task depends on the features extracted from pedestrian appearances. The best image matches are verified based on these features as identification results. The saliency information in image scenes is often exploited for feature selection in high-level vision tasks. Inspired by the pre-attentive mechanism in human visual system, we utilize the saliency information in the re-identification data to refine the person appearance in a preprocessing step. We first obtain the eye-fixation-predicting map based on the saliency analysis of image. Then this map is used to spatially weight the image features for a better appearance. Finally, we apply these processed images to the feature extraction in a standard re-identification procedure. Experiments on the widely-used VIPeR dataset show that the proposed method improves the final performance of re-identification task.

***Index Terms***— re-identification, saliency, preprocessing

## 1. INTRODUCTION

In multi-camera surveillance, it is a fundamental task to find the same persons appearing at different times and locations across cameras. But with the rapid development of today's surveillance network, it is expensive and inaccurate to identify persons in large amounts of video sequences with human efforts. Researches on person re-identification are mainly trying to solve this problem. This task is usually performed by extracting visual features and matching across candidate person samples.

The study of features and representations is an important aspect in solving the problem of re-identification, since current solutions are based on the assumption that people will not change their visual appearances rapidly during the period of surveillance. In retrospect, Gray *et al.*[1] performed viewpoint invariant pedestrian recognition using the ensemble of localized features(ELF) representation. Farenzena *et al.*[2] presented a whole pipeline with symmetry-driven accumulation of local features including the weighted HSV histogram, MSCR and RHSP descriptors. Oliver *et al.*[3] introduced the concept of Bags of Appearances to describe each person, while Liu *et al.*[4] used body structure pyramid for codebook learning and feature pooling, which is also a feature-based matching method.

On the other hand, based on the studies of visual attention in human vision system, researchers exploited the visual saliency to work as a pre-selecting and guiding mechanism for high-level cognitive tasks. Saliency detection is useful for guiding to important parts of the scene and gather detailed information in a selective way. This is applied to areas such as image classification[5] [6], object recognition[7], image segmentation[8] [9], salient object detection[10], content aware image resizing[9], etc. In the applications to high-level visual tasks, the saliency information is used as additional attentive guides, method weights and distinct feature selectors combined with existing approaches.

Recently, researchers utilized the concept of saliency to make progress in person re-identification field and achieved state-of-the-art results. This is based on an intuitive principle that people's salient appearances often help them to be more easily identified. For example, in winter, when people wear similar dark clothes, the objects carried by them or some rare-style clothing can give additional information for identifying. These usually turns to be the most salient regions in the scene. Inspired by this, Zhao *et al.*[11] first proposed a saliency matching method, exploiting the patch-level saliency distribution to match between alike persons. Further they came up with an unsupervised learning model[12] to learn and discover localized saliency to build reliable correspondence between image pairs. Afterwards, Wang *et al.*[13] proposed an unsupervised re-identification modeling approach, exploiting generative probabilistic topic models to discover salient image patches and remove background clutters. Yang *et al.*[14] proposed a color descriptor based on salient color names and utilized it to solve person re-identification problem, which also achieved a good result.

However, these models are complicated and computationally expensive. Our proposed method is different from these models in that our method is a separated process from feature extraction and successive steps. We focus on the image data themselves. Considering that parts of visual saliency mechanisms take effects at first sight, we deploy a preprocessing strategy specially for the re-id data. First, we use a pre-attentive spectral model to describe the global visual attention. Then, the saliency map is further developed to gener-

alize an eye fixation map for refining the appearance of target person in the image. Finally, we connect this process to a standard re-identification pipeline and show the improvement made by our method to the re-id performance by thorough experiments on the VIPeR dataset.

## 2. SALIENCY-BASED PREPROCESSING OF IMAGES

### 2.1. Basic Bottom-up Saliency Map

Fast, low-level spectral saliency models are able to quickly respond to visual stimulation, while they are easy to implement and combine with later processes. Here we intend to get a global map based on bottom-up saliency and deploy Achanta's frequency-tuned approach[8] to compute the basic saliency maps as the very first step. A brief formulation of Achanta's method is given as follows.

Low-level color and luminance features are exploited from CIE $Lab$ color space in three sampling channels. The saliency value $S$ of each pixel is computed as:

$$S(i,j) = \|I_u(i,j) - I(i,j)\|, \tag{1}$$

where $i$ and $j$ are pixel coordinates and $Iu$ is the mean of three channels. $I$ denotes the image blurred with a small Gaussian kernel. Parameters are kept the same with Achanta's paper[8].

However, the salient regions in this map is scattered and not smooth enough for tracking eye moves, as Figure1(b) shows. This may lose information and limit the feature selection in successive steps.

### 2.2. Eye Fixation Prediction

For the purpose of predicting the most attention-catching salient parts, we did further Gaussian filtering with a relatively big standard deviation $\sigma$. Keeping in mind that a strong blurring operation may remove too much details in the map, a dilation operation is done before that. So we get a smooth eye fixation map $E$:

$$E = DILATE(S,n) \otimes K, \quad K \sim N(\mu,\sigma), \tag{2}$$

where $K$ is a Gaussian kernel scaled to $\sum_i (K_i)$, while $\otimes$ denotes convolution operation. The dilation is applied $n$ times where we set $n$ as 11. Gaussian filtering parameter $\sigma$ is set to 20 by experience. An example resulting map is in Figure1(c), which predicts the most possible eye-catching regions in the scene.

### 2.3. Pixel-wise Preprocessing of Images

Saliency distribution can not be used directly for re-identification under viewpoint and pose varying conditions, but it influences the perception of appearance. Inspired by the weighting operation on histograms for improving the image representation, we use the previously acquired eye fixation prediction map to refine the appearance of people in image samples. As the map is in full resolution, a pixel-wise operation can be conducted.

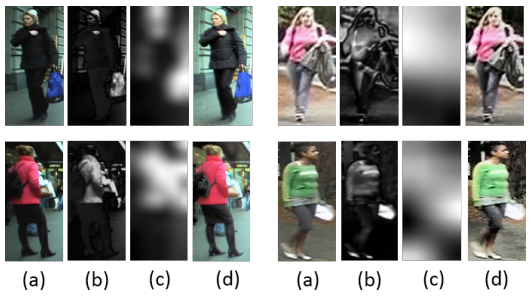


**Fig. 1**. (a) is the source image.(b) shows the raw saliency map. (c) is the generalized eye-fixation predicting map. While (d) is the resulting processed image of (a).

First, the original image is converted to HSV color space to split color and luminance channels. Then, we modify the value channel by a per-element multiplication with the map $En$, which is $L2$-normalized followed by clipping(limiting the difference between extreme values and mean value to 0.2) for better compatibility, similar to the general normalizing operation for feature vectors. The algorithm is described as follows:

$$\begin{aligned}
&1.\quad ch_k(I) = \{I_h, I_s, I_v\}, \quad k = 1,2,3 \\
&2.\quad En(i,j) = \alpha \cdot E(i,j) / \|E\|_2 \\
&3.\quad \text{if } |En(i,j) - mean(En(i,j))| > 0.2 : \\
&\qquad \text{set } En(i,j) = mean(En(i,j)) \pm 0.2 \\
&4.\quad I_v(i,j) = I_v(i,j) * En(i,j) \\
&5.\quad R = \text{ MERGE } (ch_k(I)), \quad k = 1,2,3
\end{aligned} \tag{3}$$

Where $ch_k(I)$ are channels of the image $I$ in HSV space, $I_v(i,j)$ denotes each pixel in the value channel of the image, $En(i,j)$ is the corresponding pixel in the normalized eye fixation map and $R$ is the processed result. To reduce the bias of resulting pixel values, when doing normalization, we set $\alpha$ as a scale factor so that $mean(En(i,j)) = 1$. And pixels in the final image $R$ are also re-normalized to 0-255 range. In this procedure, the hue and saturation channels of the image are kept the same. The source image and corresponding processing result is shown in Figure 1(a) and 1(d).

## 3. PERSON RE-IDENTIFICATION PIPELINE

This section briefly describes the basic re-identification procedure we adopted to evaluate the effect of the process. Since our image preprocessing mainly affects the visual appearance

of the person in image, we extract features on the modified image samples, then apply them to match persons across image sets by comparing the distances between image representations and finding the similar. The whole process can be seen as a ranking problem[15].

### 3.1. Feature Extraction

In the area of person re-identification, under low-resolution, small-size and view-changing circumstances, the color cue is the most widely used information, which is simple but efficient[16]. A basic color descriptor is the weighted color histogram. We implemented a 32-bin RGB histogram and a 30-32-bin Hue-Saturation histogram for testing in following identification steps. Each descriptor is weighted upon a $4 \times 4$ simple division of the image using weighting coefficients under a Gaussian distribution.

Moreover, we adopt a group of complicated features consisting of dense color histograms and dense SIFT sampled from overlapped patches according to [12]. Each image is segmented into a dense grid of uniform-sized local patches. From each patch, a 32-bin histogram is computed in the CIE $Lab$ colorspace with 3 levels down-sampled. SIFT features are also extracted with each patch divided into 44 cells and 8-bin quantized orientations of local gradients. All the feature vectors are $L2$-normed and then concatenated to form a 672 dimension vector($32\times3\times3+4\times4\times8\times3 = 672$). These local features are adopted by several state-of-the-art works[12] [13] [17] and the implementation is available for public. Through the experimental evaluations in the papers mentioned above, these features are shown to be more effective than many other features for person re-identification.

### 3.2. Image Matching

Now we have both simple and complex image descriptors. To achieve the goal of matching people across the image sets, a distance metric method must be adopted to measure the similarity between pairs of images. As another important aspect of research works on person re-identification, many distance metric learning methods were proposed. But in order to avoid disturbances when evaluating the effect of our method on image appearance, we adopt the Bhattacharyya distance metric instead of learning methods. By this a distance matrix is computed for matching each person in one set with its most similar figure in another set.

## 4. EXPERIMENTS AND DISCUSSIONS

We evaluate our method on the benchmark dataset VIPeR [15], which is widely used in person re-identification. We carry out a standard re-identification pipeline to show the improving effect of our method for final performance. In experiments, the salience-based processing for different feature representations is done and the results comparing with original features are shown. The combination and comparison with state-of-the-art approaches are also discussed. The performance is evaluated using the common-used Cumulative Matching Characteristics (CMC) curve, which represents the expectation of finding the correct match pair in the top $k$ matches. For convenience, we list the ranking results instead, where rank-$k$ rate is the $k$th value on CMC curve.

The VIPeR Dataset contains 632 pedestrian pairs. Each pair is made up of two images of the same individual taken from two different cameras, under different viewpoints and varying illumination conditions. All images are normalized to $128 \times 48$. Following the settings in [2] for fair comparison, we randomly sample 316 image pairs (i.e. half of the dataset) to run the experiment. All the experiments are repeated 10 times to reach a relatively stable performance. The average ranking results using RGB histogram, H-S histogram, dense feature(as in[12]) and the corresponding pre-processed versions are listed below in Table 1. Where features based on our processed appearance are denoted by *-Pre. While DF is the abbreviation for the Dense Feature in [12].

| Rank | R-1 | R-5 | R-10 | R-15 | R-20 |
|---|---|---|---|---|---|
| RGB Hist | 4.19 | 9.40 | 13.51 | 17.69 | 21.34 |
| RGB-Pre | 4.00 | 9.02 | 13.90 | 17.65 | 20.79 |
| H-S Hist | 11.07 | 25.48 | 34.24 | 40.52 | 46.68 |
| HS-Pre | 11.38 | 25.38 | 35.46 | 42.53 | 48.52 |
| DF[12] | 9.24 | 21.77 | 30.54 | 37.94 | 43.45 |
| DF-Pre | 10.06 | 23.58 | 32.72 | 38.86 | 44.62 |

**Table 1**. Test on half VIPeR

From the results we can see that our refining method based on eye-fixation-predicting map is effective for some cases and improved the re-identification result, but not always so. As for RGB histogram, our method fails. This might be caused by the compatibility of refining method with RGB feature.

In addition, we did further experiment on the full VIPeR set(i.e. all 632 image pairs). The result is listed in Table 2. In these experiments, when sample images are pre-processed by eye-fixation maps, the extracted features all lead to a better result. Because the full VIPeR set were used, there's no need to do random resampling and the repeating results are all the same. Perhaps there still exists bias in the random selection of 316 image pairs, which leads to the failure in previous RGB-histogram tests.

To further evaluate the proposed method, we combine it with state-of-the-art re-identification approaches. The S-DALF method[2] exploits three kinds of features and weights them by exploiting symmetry perceptual principles. We process the image samples with generated eye fixation maps by our method, then combine them with SDALF. Table 3 shows the comparison, where basic results in the first line are got by single-shot SDALF. The baseline result comes from the

| Rank | R-1 | R-5 | R-10 | R-15 | R-20 |
|---|---|---|---|---|---|
| RGB Hist | 2.85 | 7.12 | 9.18 | 11.08 | 13.61 |
| RGB-Pre | 3.32 | 6.33 | 9.02 | 11.71 | 13.77 |
| H-S Hist | 7.75 | 19.15 | 25.16 | 30.54 | 34.34 |
| HS-Pre | 8.07 | 18.67 | 25.63 | 31.49 | 35.92 |
| DF[12] | 6.49 | 15.35 | 23.42 | 28.01 | 31.49 |
| DF-Pre | 7.44 | 15.66 | 23.42 | 28.01 | 32.59 |

**Table 2**. Test on full VIPeR

source code published by the author of [2]. But the experiment data we get using the code is slightly different from the results published in the origin paper. This is perhaps because of the small differences existing in experimental environment and not vital for the comparison.

| Rank | R-1 | R-5 | R-10 | R-15 | R-20 |
|---|---|---|---|---|---|
| original | 18.35 | 38.67 | 49.65 | 57.06 | 63.61 |
| preprocessed | 18.70 | 39.46 | 50.63 | 57.37 | 63.10 |

**Table 3**. Test for combining with SDALF on VIPeR

From the result we can see that our method showed overall positive effects when combining with the SDALF method. Meanwhile, we can observe that at R-20 the new method with our process performs not as well as original method. However we know the fact that, denoting the ability of finding the best matching pairs at first time, the R-1 metric is more important than the latter ranks. The metric importance is decreasing with the rankings increase, this is why the ranks after 20 are often omitted.

To sum up, in these experiments on VIPeR dataset, when sample images are pre-processed by eye-fixation maps, the extracted features lead to a better result in most of the cases. And experiments showed that for the combined features that are complex enough, our method also has a positive effect.

## 5. CONCLUSION

In this work, we propose a preprocessing strategy for sample images in person re-identification task. The method is based on the eye-fixation map developed from a bottom-up saliency method, simulating the early processing in visual attention mechanism. Experiments show that it can practically improve the performance of a standard re-identification task, with some failure cases discussed. This method is easy to implement and has little time cost without any need for learning procedure. This process is also easy to combine with other methods as it only affects the preprocessing step. Future work can be done to test the performance when it is added to more other approaches including feature-learning methods and metric learning methods.

## 6. ACKNOWLEDGEMENTS

This work is supported by the NSFC 61273274, 61370127 and 61201158, NSFB4123104, FRFCU 2014JBZ004, Z131110001913143, and the Fundamental Research Funds for the Central Universities (2015YJS047).

## 7. REFERENCES

[1] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV 2008*. 2008, vol. 5302, pp. 262–275, Springer.

[2] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*. IEEE, 2010, pp. 2360–2367.

[3] Javier Oliver, Antonio Albiol, and Jose Manuel Mossi, "Re-identifying people in the wild," in *ICASSP*. IEEE, 2013, pp. 2302–2306.

[4] Hong Liu, Liqian Ma, and Can Wang, "Body-structure based feature representation for person re-identification," in *ICASSP*. IEEE, 2015, pp. 1389–1393.

[5] Yongzhen Huang, Kaiqi Huang, Yinan Yu, and Tieniu Tan, "Salient coding for image classification," in *CVPR*, 2011, pp. 1753–1760.

[6] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *CVPR*, 2012, pp. 3506–3513.

[7] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and I. W. H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 769–779, 2014.

[8] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.

[9] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu, "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.

[10] Jianming Zhang and Stan Sclaroff, "Saliency detection: a boolean map approach," in *ICCV*. IEEE, 2013, pp. 153–160.

[11] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *ICCV*. 2013, pp. 2528–2535, IEEE.

[12] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *CVPR*. 2013, pp. 3586–3593, IEEE.

[13] Hanxiao Wang, Shaogang Gong, and Tao Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *British Machive Vision Conference*, 2014, BMVC 2014.

[14] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li, "Salient color names for person re-identification," in *ECCV*. 2014, ECCV, pp. 536–551, Springer.

[15] Doug Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," *IEEE International Workshop on Performance Evaluation for Tracking & Surveillance Rio De Janeiro*, 2007.

[16] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, *Person re-identification*, vol. 1, Springer, 2014.

[17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159.