# Specific Patches Decorrelation Channel Feature on Pedestrian Detection

Xue-ming Ding and Dong-fei Ji
School of Optical Electrical and Computer Engineering, University of Shanghai for Science and Technology, China

**Abstract:** *Typical Local Decorrelation Channel Feature (LDCF) for pedestrian detection generates filters derived from decorrelation for each entire positive sample, using Principle Component Analysis (PCA) method. Meanwhile, extensive pedestrian detection methods, which utilize statistic human shape to guide filters design, point out that the head-shoulder area is the most discriminative patches in typical classification stage. Inspired by above mentioned local decorrelation operation and discriminative areas that most classifiers indicate, in this paper we propose to integrate human shape priority into image patch decorrelation to generate novel filters. To be specific, we extract covariance from salient patches that contain discriminative features, instead of each entire positive sample. Furthermore, we also propose to share covariance matrix within grouping channels. Our method is efficient as it avoids extracting uninformative filters from redundant covariance of convergent patches, due to embedded prior human shape info. Experiments on INRIA and Caltech-USA public pedestrian dataset has been done to demonstrate effectiveness of our proposed methods. The result shows that our proposed method could decrease log-average miss rate with detection speed retained compared to LDCF and most non-deep methods.*

**Keywords:** *Specific patches partition, decorrelation, shared covariance, channel features, average human shape.*

*Received May 22, 2019; accepted May 28, 2020*
*https://doi.org/10.34028/18/4/1*

## 1. Introduction

As [20] has raised, there remains a remarkable gap between human and machine to exploit in pedestrian detection. Current pedestrian detection researches mainly focus on following directions.

The first part which is extensively discussed is the feature extraction stage. Nam *et al*. [9] studied image channel feature transformation and put forward aggregated channel feature utilizing totally ten channels. Zhang *et al*. [17] proposed informed haar-like features using human template to generate hand-designed filters. Shen *et al*. [12] modified traditional Local Binary Pattern (LBP) method and proposed nearby differential statistic feature. Filters can also function as projection. Zhao *et al*. [21] argued random projection features with shape prior improves pedestrian detection. The second important issues been studied recently is the classification module. Particularly popular classification methods could be categorized into Support Vector Machine (SVM) and AdaBoost methods along with feature extraction in pedestrian detection. Ohn-Bar and Trivedi [10] proposed dual-stage group cost-sensitive real-boost routine as an improvement of traditional boosting classifier. Baek *et al*. [2] put forward additive kernel SVM with cascade structure.

The last but not the least direction of pedestrian detection being extensively exploited is the deep learning based methods. Tian *et al*. [14] and Zhou *et al*. [22] use deep networks, like Convolutional Neural Network (CNN), to extract higher level features as number of lays increase. Ouyang *et al*. [11] integrated feature extraction and classification stage in traditional pedestrian detection approach into a deep model and notably improved the final detection performance. For more application scenarios, Al-Najdawi *et al*. [1] propose a kind of pedestrian tracking system based on Kanade-Lucas-Tomasi (KLT) features.

For some scenarios where deep learning models cannot be deployed, we focus on the first issue called traditional manual feature extraction methods. Motivated by the decorrelation operation in Local Decorrelation Channel Feature (LDCF) [9] which utilizes a shared covariance coefficient across each entire image patch, we integrate patches based decorrelation operation into feature extraction. Meanwhile, inspired by weight map generated by node weight in decision trees, we propose a patch selection method based on average human shape constancy. Our method focuses on patch that is quite discriminative, i.e., head shoulder area whose feature has much higher weight in above mentioned classification stage. These patches could generate covariances that reveal feature of corresponding patches more precisely. What's more, we share the extracted filters within grouping channels.

Motivated by connection between transformation in LDCF and filters in [18] has proposed, we have made following contributions:

- We propose to extract fixed covariance matrix from specific local patches instead of entire image.
- We incorporate human shape and statistical weight map as prior to assign specific patches that the shared covariance matrix will be learned from. Furthermore, we also propose to share covariance matrix within grouping channels.
- Numerous experiments have been conducted to verify that our method could achieve a balance between detection speed and accuracy.

Our paper is structured as follows: section 2 reviews relevant researches related to decorrelation on pedestrian detection and features based on statistic human shape; section 3 devotes to analyse the rationality and feasibility of our proposed method and introduce the decorrelation features of specific patches within and cross the channels, finally propose the specific patches decorrelation channel feature within grouping channels; section 4 displays experiment results and compares to state of the art; section 5 is summary of the paper.

## 2. Related Work

All above mentioned methods in section 1, such as random projection methods and filtered channel feature methods as they named, could be interpreted as a post-processing along with extracting normal channel features. Here are two main methods same as above, one mainly uses the decorrelation operation, and the other integrates the human shape prior information to further improve the pedestrian detection performance.

### 2.1. LDCF

Based on [3, 5], LDCF [9] appends decorrelation operation to ten transformed channels derived from original image. To be specific, they prove that detector integrating local decorrelated data into simple orthogonal trees performs as accurate as combination of oblique split and correlated data does. To avoid bloated computation cost in oblique classifier, they preferred the former combination. The typical procedure of LDCF is shown in Figure 1. Patch P with size $m \times n$ denotes sampling patch in a $w \times h$ detection window. Averaged Σ is fixed covariance matrix shared across all channels, it's learned from a collection of cropped pedestrian samples. Using eigen decomposition, Σ could be decomposed as $\Sigma = Q \wedge Q^T$ where $Q$ is orthogonal matrix and $\wedge$ is a diagonal matrix of eigenvalues. Thus $Q^TP$ denotes decorrelation operation for patch P. For detection window with size $m \times m$, corresponding covariance matrix has size $m^2 \times m^2$, which is identical to size of $Q$. Computation of $Q^TP$ may be time cost. The decorrelation computation could be further simplified by conducting a series of convolution operations. By selecting k eigenvectors in orthogonal matrix $Q$ according to top k maximum eigenvalues using the Principle Component Analysis (PCA), the derived $Q'$ has size of m $m \times m \times k$. Then the tedious computation of $Q^TP$ could be simplified by convolution operation between k filters ($Q'$) and the detection images. The output will be treated as final feature of the image and will be sent to the classifier.

What's more, [9] argued that LDCF will achieve state of the art performance by tuning parameters, including namely size of training dataset, sliding window stride, sliding window size and number of bootstrapping stages.

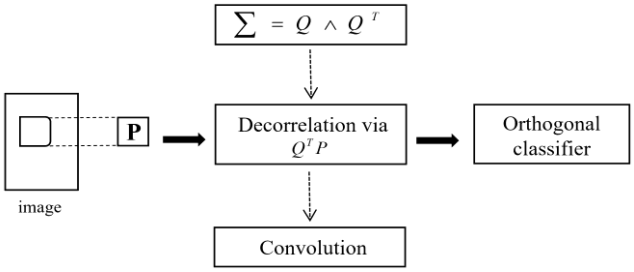



Figure 1. Calculation in LDCF: horizontal arrows indicate main procedure of LDCF, vertical arrows indicate two acceleration of LDCF.

### 2.2. Statistic Human Shape Features

Zhang *et al*. [17] proposes efficient Informed Haar-like Feature (IHF), which is hand designed filter utilizing human shape. IHF uses 1 and -1 as weight values for filters. Shen *et al*. [13] modifies IHF and proposes coarse and fine granularity shape statistic distribution feature, called Coarse-granularity Shape Statistics' Distribution feature (CSSD) and Fine-granularity Shape Statistics' Distribution feature (FSSD). They divide filters in IHF method into fine blocks with specific weight each. In detail, the weight maps are learned from statistical human shape info using imbalance embedding Linear Discriminative Analysis (LDA). The filters become complicate and capable to distill local details of pedestrian. By utilizing statistic information of human shape, the CSSD and FSSD have made improvement on pedestrian detection. Different from IHF which is categorized as nearby feature, Cao *et al*. [4] proposes no-nearby features named Side-Inner Difference Feature (SIDF) and Symmetrical Similarity Features (SSF) inspired by appearance constancy and shape symmetry, as is illustrated in Figure 2. In brief, SIDF could be written as $f(A,B) = S_A / N_A - S_B / N_B$, where $S_A$ and $S_B$ are the sum of pixels in A and B in each channel, $N_A$ and $N_B$ are the number of patches in $A$ and $B$. SSF could be written as $f(A,A') = |f_A - f_{A'}|$, where $f_A$ and $f_{A'}$ signify the features of $A$ and $A'$ respectively.

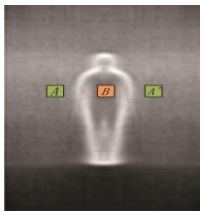


Figure 2. Human shape symmetry and appearance constancy.

CSSD and FSSD features divide the Harr wavelet template into more precise regions with weights by introducing the prior information in the average humanoid template, and obtain higher recognition accuracy. Based on the symmetry of the average human shape template, SIDF and SSF features select different regions to calculate the different side difference features and symmetrical similar features, and get better detection results. Because of considering the correct prior information, the pedestrian detection accuracy is improved. LDCF algorithm extracts the covariance of all areas of positive samples without considering any prior information, so the decorrelation operation will ignore those areas with more discriminative power, that is, the decorrelation is not complete. Therefore, the statistical human shape prior information can be integrated into the decorrelation operation to obtain better detection results.

## 3. Specific Patches Decorrelation Channel Feature

Decorrelating the feature channels of the input image can improve the classification performance of the orthogonal classifier, such as gradient lifting decision tree. The LDCF decorrelation algorithm extracts covariance from all the random samples of positive samples, and ignores some specific areas with high identification to some extent. The proposed algorithm extracts the covariance matrix from the specific scattered areas, and the generated filter shares with the whole detection window within grouping channel. Figure 3 shows the general scheme of our proposed method.

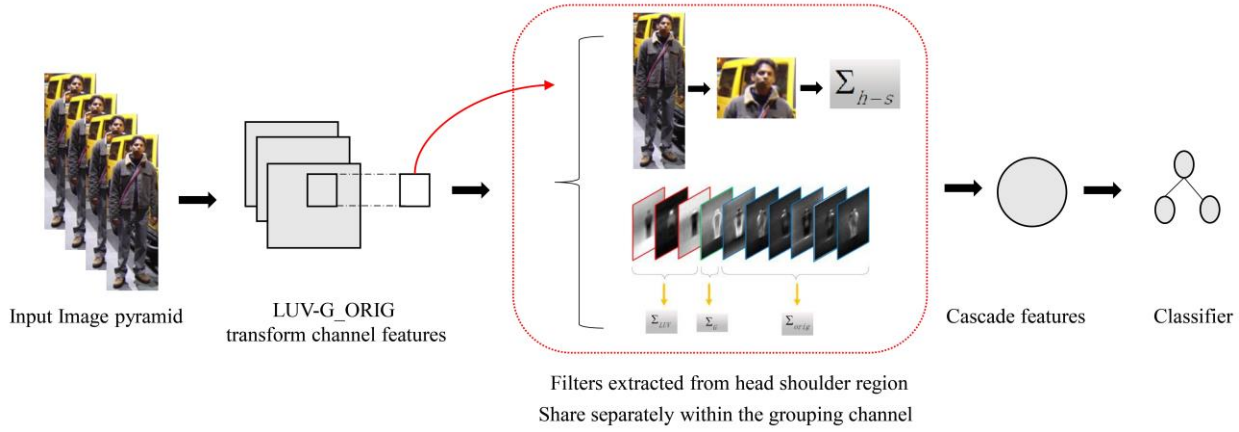



Figure 3. General scheme of the proposed method.

### 3.1. Feasibility of Specific Patches Decorrelation

In order to predigest computation in LDA, which generates quite effective detector following correlated data, LDCF proposes to share fixed covariance across all image patches. It not only comes to the expectation in speed, but also benefits data regularization. Nevertheless, sharing fixed covariance across all patches each channel may have defect that learned filters have weak generation ability, because negative samples exceed positive samples in quantity, which could be observed in the public datasets.

Based on above analysis, we propose to integrate statistic human information into decorrelation operation. Our innovation could be supported by two reasons. Its rationality is due to the observation that it is feasible to capture local features from partial average human shape through above mentioned decorrelation operation. The reason is that the human shape was averaged from cropped pedestrian images with most patches accurately reflecting actual position of corresponding patch. Therefore, partial regions could also reveal partial features. The feasibility could also be verified. For purpose of acceleration, LDCF uses top k eigenvectors to conduct convolution to simplify computation of $Q^T P$ which works as decorrelation on patch P. $Q \wedge Q^T$ is eigende composition for covariance learned from patches across the whole image. Obviously, $Q' \wedge (Q')^T$ could also be regarded as the decorrelation for covariance learned from partial patches. Therefore, decorrelation for local patches could also be simplified by means of convolution.

Patches separation topology will be elaborated below. Generally, the type of partition could be categorized as within-channel level and cross-channel level. The first within channel partition method sets could be seen as a division within each single channel according to human shape prior. Precisely, divided patches are achieved according to response of channel transformation. We also conduct experiments on equally divided blocks on finely cropped pedestrian

sample images. In addition, the cross channel level partition methods consists of average covariance that is learned from multi channels. We share it across the channels it is learned from as well.

### 3.2. Patches Partition within the Channel

In the classification stage, the most discriminative area is the head shoulder area, as shown in Figure 12 through the weight map calculated by the decision tree in the classification stage. Different colors represent the weight value of the corresponding area. The brighter the color in the color bar, the higher the weight value. Therefore, extracting covariance from head shoulder region of positive samples instead of all regions of positive samples in traditional LDCF can get better detection effect. Denoting the size of patches that share identical covariance matrix as $S_P$ with size $m \times n$, it is restricted to size of detection window $w \times h$, i.e., $m \le w$, $n \le h$. It's lower bound is intuitively set to be size of head area in average human shape map. Because it can be seen from the weight map above, the head area is most discriminative patch for classifier and can't be cropped. Denoting $\sum_i$ as $i_{th}$ learned average covariance matrix, its corresponding filter would be $filter(i)$. We are going to generate four 5×5 filters each channel.

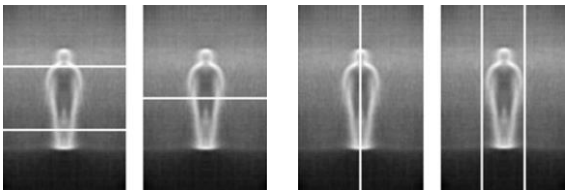
The learned filters in within channel partition method sets could be denoted as:

$$filter(i) = [F_1^i \| F_2^i \| \ldots \| F_j^i], \; i = 1,2,...10, \; j = 1,2,... \quad (1)$$

Where $filter(i)$ indicates filters for channel $i$ and $F_j^i$ indicates $j_{th}$ filter in a specific patch in channel $i$. The final output of the filters will be concatenated and vectorized as entry of the boosting classifiers.

Figure 12 is on behalf of the response diagram of feature channel in different methods. According to the response size of all the channels, the region with the highest response degree is selected to extract the covariance matrix. As you can see, all the feature channels contain the head shoulder region. More pedestrian information can be taken into account to improve the detection effect by decorrelating the high response area as a prior information. Next, several region segmentation methods are proposed for contrast experiment. The covariance matrix used to generate the filter will be extracted from these specific regions by random sampling.

- Horizontal and vertical partition. Except for human shape based local partition method, we further carry on random partitions on positive sample training images. Recall that the head-shoulder area is most discriminative area proved by the weight map of boosting classifiers, we intuitively divide the cropped image into several blocks, which equals each other in size, as depicted in Figure 4. LDCF [9] argued that effective filters could be learned via reshaping top 4 eigenvectors in decorrelation, thus making 4 filters per channel. The PCA Foreground [18] method learns 4 filters from both positive and negative samples respectively, making 8 filters per channel in total. To reduce time cost and improve detection performance at the same time, we select top 2 eigenvectors when images are divided equally into 3 or 2 parts, making final detector with 6 or 4 filters for each channel.



a) Horizontal partition. b) Vertical partition.

Figure 4. Equally partition on positive sample.

- Crop on salient human contour. Obviously, average covariance that reflects correlation of pedestrian information must be learned from patches solely covering pedestrian samples. Annotated pedestrian samples contain much non-pedestrian regions, making it hard to distill covariance purely revealing pedestrian. As a consequence, for sake of capturing salient features on pedestrian, we use cropped average human shape map to obtain more accurate pedestrian samples. To be specific, we make a fine crop operation once again based on original annotated pedestrian samples. The sketch map is shown in Figure 5. Red box indicates patches that the covariance matrix would be derived from. Size of learned filters are identical to that in LDCF which is 5×5×4 per channel.

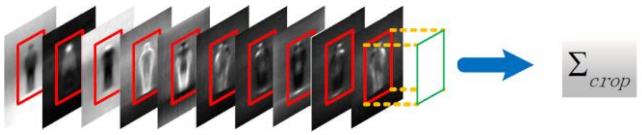



Figure 5. Crop on human countour.

- Crop on head-shoulder area. According to most significant area (head-shoulder area) as the weight map has illustrated, we intuitively crop average head-shoulder area in pedestrian samples where the new covariance will derive from. The sketch map is shown in Figure 6.

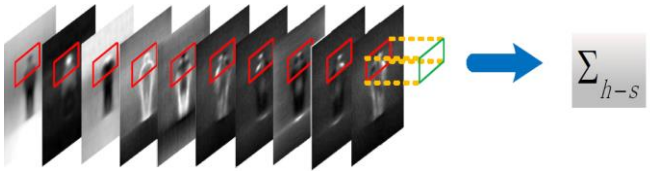



Figure 6. Crop on head-shoulder area.

### 3.3. Patches Partition Cross The Channel

In comparison with cropping within channel images

into blocks for derivation of covariance, we also propose to conduct division operation on channel level, namely one or more channels will share fixed covariance matrix. In cross channel partition method sets, the final learned filters are depicted as:

$$filter(I_i) = F_{I_i},\ I = [1, 2, ..., j] \quad (2)$$

Where vector $I_i$ indicates $i_{th}$ set of channel that will share fixed covariance and $j \leq 10$, $F_{I_i}$ means filters correspond to $I_i$. For example, if nearby channels are set to share common fixed covariance, then $I_1$=[1, 2], $I_2$=[3, 4] …

Covariance averaged from all channels. On the observation that channel feature transformation of positive samples may bring about covariances that relate to each other, we integrate covariances derived from ten channels of each cropped pedestrian sample, which is shown in Figure 7. The united covariance will be decomposed to generate common filters which will be applied on all channels.

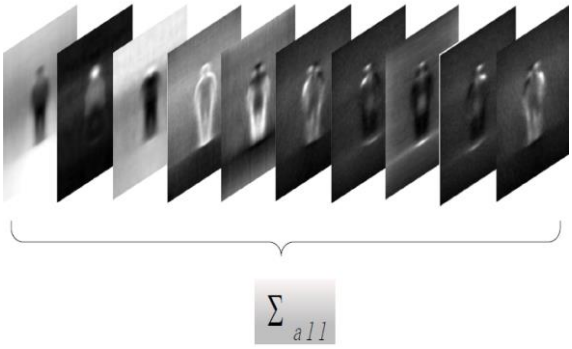



Figure 7. Covariance learned from all channels.

- Covariance shared within nearby channels. We also conduct experiments on averaging covariances combinations of nearby channels, which on behalf of random assemblies and function as comparison experiment. The sketch map is shown in Figure 8.

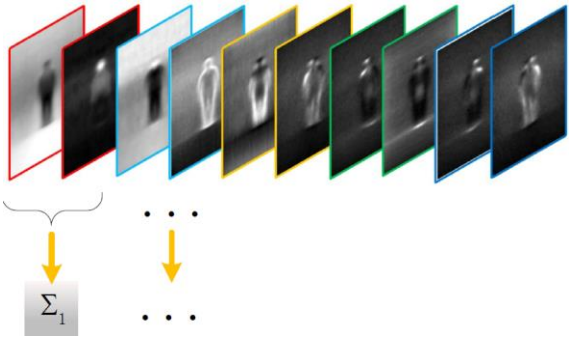



Figure 8. Covariance learned from nearby channels.

Covariance shared within grouping channels. We divide 10 channels into three groups: LUV channels, the gradient magnitude channel and six orientation channels which is abbreviated as LUV-G-ORIG below, which is shown in Figure 9. The intuition is clear: transformation of channels in each division group may correlate with each other. Each group of channels will share fixed covariance matrix.

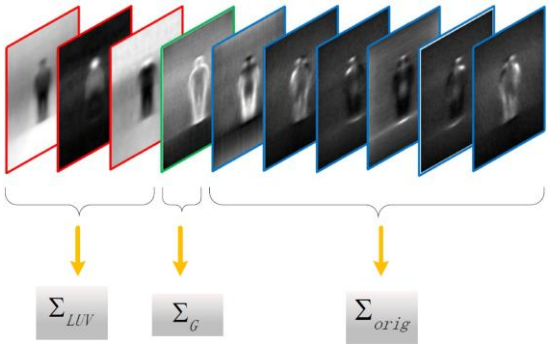



Figure 9. Covariance learned from specific channels.

### 3.4. Covariance Matrix Analysis

The covariance matrix could be written as $\sum_{(x,y),(x',y')} = C(x'-x, y'-y)$ following [9]. The covariance computation procedure for patch P is depicted as Figure 10. For patch P with size 5×5 in sliding window detection, its learned covariance matrix has size 25×25.

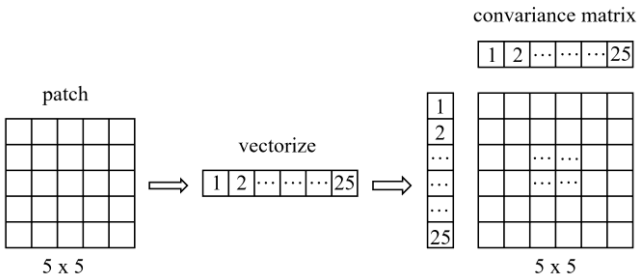



Figure 10. Covariance computation scheme.

The head shoulder area is the most discriminative area for pedestrian expression, because pedestrian occlusion often occurs in the lower half of the body, and the head is rarely occluded, so the content of head information packet is larger. By extracting the covariance of specific area (head and shoulder), we can get more discriminative filters. PCA uses maximum projection variance to reduce dimensions, that is, to select feature projection dimensions with more information. The more information it contains, the greater the variance, so the optimization goal is to maximize the projection variance. By solving the characteristic variance after coordinate transformation and using Lagrangian duality, it is found that the maximum projection variance is the characteristic value of the maximum covariance matrix.

The essence of decorrelation is to reduce the dimension of Aggregated Channel Features (ACF) features and remove the correlation of features after matrix multiplication. Therefore, the following principles should be followed in decorrelation:

1. The larger the projection variance of the feature in each dimension, the better.

2. The smaller the correlation between different dimensions, the better.

For the orthogonal decision tree model, it is likely to affect the final classification effect by training the expression features with high correlation. According to the knowledge of linear algebra, the covariance of feature itself represents its variance, and the covariance between different features represents its correlation. Therefore, the core of decorrelation operation is to extract the covariance matrix. The optimization objective can be expressed by the covariance matrix. The ideal covariance matrix is as follows:

$$\begin{bmatrix} \delta_{11} & 0 & ... & 0 \\ 0 & \delta_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \delta_{kk} \end{bmatrix} \quad (3)$$

Figure 11 shows the visualization of the covariance of head shoulder region based on the intra channel region segmentation method and intra group based on the cross channel region segmentation method. Take the area p with size as an example, the leftmost color bar represents the value of the corresponding covariance matrix. The extracted covariance matrix should make the extracted features meet the maximum variance, that is to say, the diagonal feature value is the largest, and the correlation between different features is smaller. Thus, the correlation between different features is 0 in the best case, such as the non-diagonal element in the ideal covariance matrix above. After decorrelation of ACF features, more information should be included in the reduced dimension features, and the larger the trace of covariance matrix, the better. It is proved that this method contains more information and the operation of decorrelation is more thorough. It can be seen that the graph consists of 25 small matrices. Each small matrix represents a covariance matrix with other pixels in the region. As you can see, the adjacent areas are highly correlated. As the region expands from the center to the outside, the corresponding covariance difference will also decrease, and the regional correlation will decrease.

When the center is expanded outwards, the value of the covariance matrix will decrease to some extent, that is, the resolution of the covariance matrix will decrease with the expansion of the region. The smaller the region P is, the more relevant information will be contained in the extracted covariance matrix, making the decorrelation operation more effective, which will be helpful for the extraction of pedestrian expression features. This prior information should be taken into account when adopt patches partition within the channel. Therefore, we can speculate that extracting the covariance matrix from the head shoulder region, which is the highest weight in the classification stage, can obtain more discriminative expression features, thus obtaining higher detection accuracy.

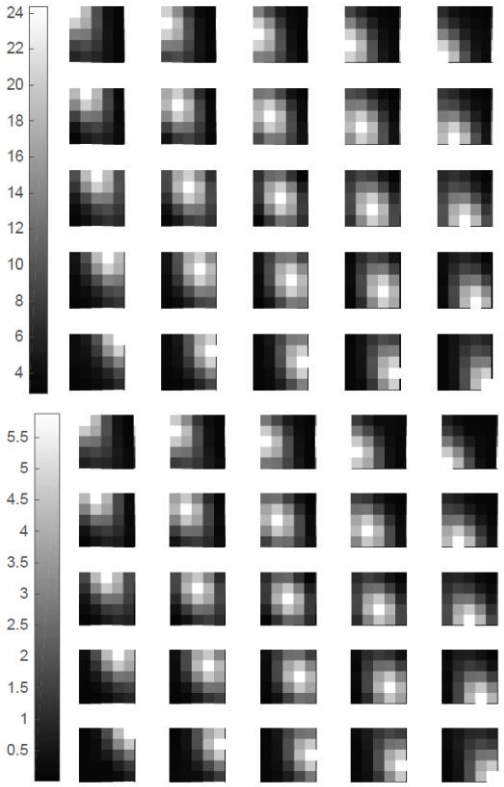



Figure 11. Visualization of averaged covariance matrix on channel L, upper: within channel partition method, lower: cross channel partition method, which are similar but not identical.

## 4. Experiments and Discussions

### 4.1. Common Experiment Settings

In order to verify the effectiveness of our proposed method, we have conduct massive experiments on INRIA and Caltech-USA pedestrian datasets. Total number of positive samples is about 24740 for the INRIA dataset and 24498 for Caltech-USA dataset. Our experiments settings on Caltech-USA dataset will refer to [9]. The detailed settings are depicted below. Size of the detection window is set to be 128×64, and the positive samples is cropped from the annotated images. Each annotation of pedestrian follows a rollover operation to relieve the misalignment phenomenon. We incorporate a down sample method to shrink original channel maps, i.e., from size 128×64×10 to size 64×32×10. The size of learned filters from shared covariance is 5×5×4 each channel.

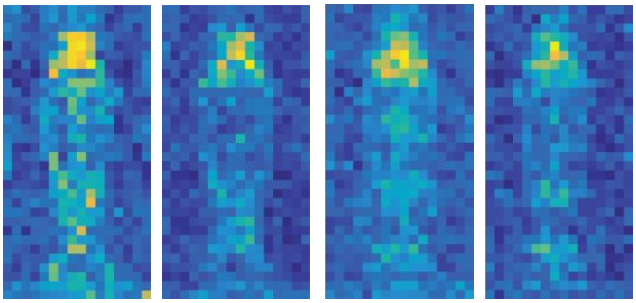
We make use of AdaBoost algorithm to carry out feature selection, as it is simple and efficient that integrates decision trees whose depth is set to 3 to make up a strong classifier. The AdaBoost procedure is conducted by rounds (at most 4096 weak classifiers) trained with 10,000 negatives which are bootstrapped from negative samples. Also, we utilize public toolbox [6] and evaluation tools [6] for detector generation and evaluation.

- Performance measurement. As is widely accepted, the log-average Miss Rate (MR) versus False

Positives Per Image (FPPI) plot is effective to measure performance detectors. To identify accurate detection result, we traditionally use the overlapping ratio as $P(B_{dt}, B_{gt}) = (B_{dt} \cap B_{gt})/(B_{dt} \cup B_{gt})$ to describe the overlapping between detection truth and ground truth, where $B_{dt}$ and $B_{gt}$ separately denote the detection bounding box and the ground truth bounding box. We set threshold of P as common used 50%, namely if $P(B_{dt},B_{gt})$>50%, corresponding detection bounding box of localization is recognized as accurate.

## 4.2. Comparisons with the Partition Methods

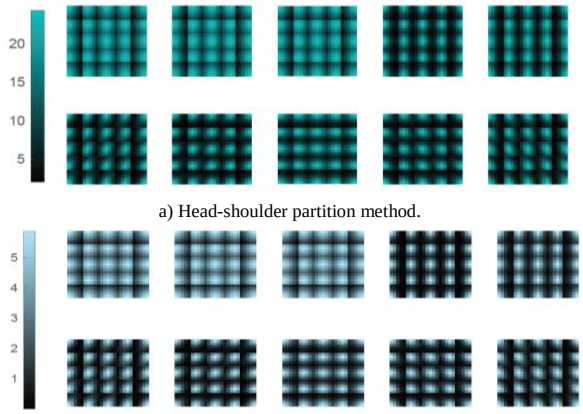
In the general scheme of our method and related methods, such as LDCF and Rotated filters, features are extracted from image samples and work as entry of classifiers afterwards. Feature weight map denotes weight of features in node of the boost decision trees. Accordingly, higher weight one feature has, more discriminative its corresponding patch is. To some extent, the discrimination ability of detectors could be revealed in the learned weight map. We are going to compare weight maps generated by proposed methods with that of other methods. The weight maps shown in Figure 12 are all get from corresponding classifiers trained on INRIA dataset. Size of weight map is set to be 32×16 in these methods. We choose two of our proposed methods for comparison, which are head-shoulder partition method in within channel partition method sets and the LUV-G-ORIG partition method which belongs to cross channel partition method sets, for they achieve much better performance in each category, respectively. As can be seen, the head region has much higher response in all above methods generally. Our detector depicts the human head area more clearly and distinctly than other methods. In detail, our proposed method assigns much larger weight to the head-shoulder area. The result is that we could decorrelate detection regions using covariance learned from original head area, making the final detector capable of capturing salient features.



a) Within ch. b) Cross ch. c) LDCF. d) Checkboards.

Figure 12. Weight maps learned from tree-based classier, lighter color indicates larger weight, therefore our method could capture more salient features in the head area in comparison with other methods (ch. denotes the channel).

The learned covariance matrixes of head-shoulder partition method and LDCF are displayed in Figure 13. With same covariance size, our method improves the detection performance with speed retained. It can be observed that the within channel partition method sets and cross channel partition method sets all perform better than typical LDCF from Tables 1 and 2. However, the former slightly outperform the latter for reason that the former method capture more discriminative human shape information of local region. In within channel partition method sets, the head-shoulder partition achieve MR of 12.26%, which is much lower than others. It could be concluded that the discrimination ability of local patches is not in proportion to their size. Such a phenomenon verifies our conclusion as before that the head area is much more discriminative than other areas.




a) Head-shoulder partition method.

b) LDCF.

Figure 13. Visualization of learned covariance over ten channels.

When dividing average human shape into parts, the horizontal division methods could capture distinct features. As could be inferred that the vertical division method may fail capturing informative features about pedestrians. This is clear as occlusion mostly occurs in lower parts of human shape and the vertical division method fails to decorrelate head regions. It also verifies the observation that the head-shoulder area contains much more salient features in practice. LUC-G-ORIG grouping channel method can get the best performance in the cross-channel segmentation method sets, so we directly choose the head shoulder region to extract covariance and share it within the grouping channel. The results are shown in the Figure 14, 15, and 16.

Table 1. Within channel partition method performance on INRIA dataset.

| Within channel partition method set | Log-average MR(%) |
|---|---|
| Head | 12.30 |
| Human contour | 13.35 |
| Head-Shoulder | 12.26 |
| All croped | 13.35 |
| Horizontal division-3 | 13.19 |
| Horizontal division-2 | 12.31 |
| Vertical division-3 | 14.77 |
| Vertical division-2 | 15.29 |

Table 2. Cross channel partition method performance on INRIA dataset.

| Cross channel partition method set | Log-average MR(%) |
|---|---|
| All in one | 13.31 |
| LUV-G-ORIG | 13.30 |
| Neighbour combination | 13.38 |

## 4.3. Comparison with State of the Art

- Evaluation on INRIA dataset. The INRIA public pedestrian dataset includes training data with 2416 human annotations in 614 images and testing data with 1132 human annotations in 288 images. Total images of training data and testing data are 1832 and 741 respectively.

The deep learning method, such as F-DNN [7], RPN+BF [16], which utilize GPUs to operate computations for deep layers may fail in real-time use. For some scenarios, deep learning models cannot be deployed. Therefore, we will mainly focus on comparisons between our methods and other non-deep pedestrian detection methods. Non-deep pedestrian detection methods also achieve state of the art performance, such as LDCF [9], IHF [17], Sketch Tokens [8] and Rotated Filters [19]. Our method could be seen as modified version of LDCF. Our comparison also includes the baseline methods VJ [19] and so on.

Following typical scheme in LDCF, we utilize four stages to train the classifier. The total negatives would be 20000 with incorporation of hard negative mining strategy. The positive images will be shrink to size 32× 16 as entry vectors of the classifier. The sliding window size is 64×32 and the sliding stride is set to be 4. We use AdaBoost to integrate decision trees with depth 3 as classifier. The experiment results are displayed in Figure 14, in which the compose of head-shoulder based partition method and sharing within grouping channels method is abbreviated as 'ours'. It could be observed that our proposed method outperforms most state of the art. In detail, we outperform LDCF with a 1.56% decrease in MR. Such a result verifies that sharing learned covariance across specific patches guided by statistical average human shape could improve pedestrian detection.

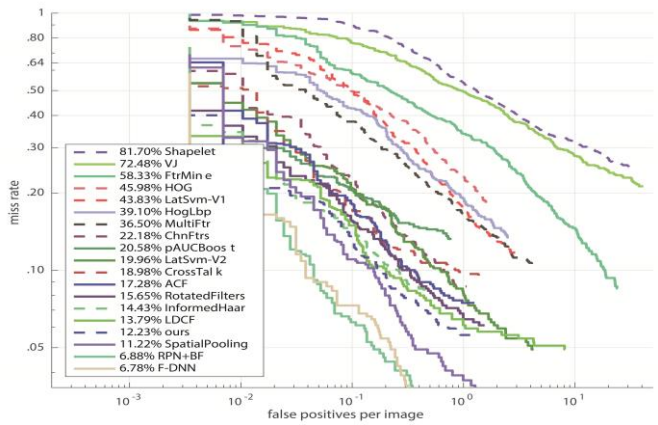



Figure 14. Overall result on INRIA dataset.

- Evaluation on Caltech-USA dataset. The Caltech public pedestrian dataset could be categorized into training data (set00-set05) and testing data (set06-set10) which are all obtained from a video. The training data has 4250 images with 6325 pedestrian annotations and the testing data has 4024 images with 5051 pedestrian annotations. We will follow the optimal settings [9] has proposed, i.e., training data will be obtained every $3_{th}$ frame of original video, the sliding window is set to be 120×60 and sliding stride is 2. Also, we utilize 5 bootstrapping stages to train our classifier. We are going to compare our method with state of the art methods on Caltech-USA dataset. The experiment results are displayed in Figure 15.

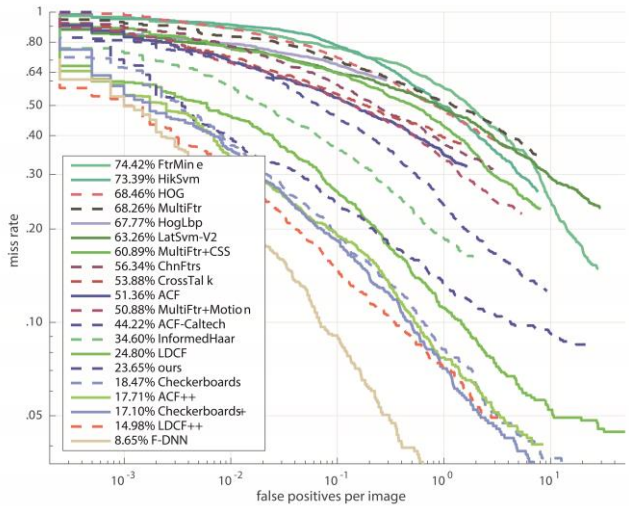



Figure 15. Overall result on Caltech-USA dataset.

Our proposed method not only simplifies computation in training stage comparing to LDCF, but also reduces the MR to some extent. More specifically, our method reduces 1% in MR, with the MR of LDCF and our method being 24.8% and 23.28% respectively. The LDCF based boosting methods achieve much better performance, such as LDCF++ [15]. Nevertheless, they are generally data-hungry thus rely on data augmentation. It cannot be denied that with data augmentation, our method could contribute to a significant promotion in detection performance as well. Such an observation will also be testified dealing with extension of additional features, such as optical flow which is incorporated in checkerboards+ [16] method. In addition, we prevail the deep learning based methods in run time as our method could be execute on only one CPU with single core, which benefits to real time application.

We also made evaluations under partial occlusion (1-35% occluded) conditions on Caltech-USA pedestrian test dataset. Figure 16 shows the evaluation results. Obviously, the detection performance of all methods that has been tested weakens sharply as the occlusion occurs. Yet, our method also outperforms typical LDCF under partial occlusion conditions. We

conclude that head region based partition method does good to discriminative features extraction as it focuses on the head area, thus is robust to occlusions as occlusion conditions mainly occur in lower parts of human shape in Caltech-USA dataset.

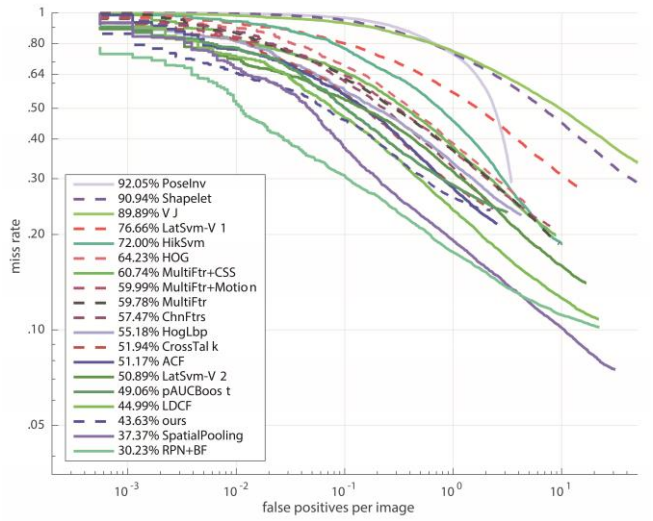



Figure 16. Detection performance under partial occlusion.

### 4.4. Running Time Analysis

Our detector is implemented with MATLAB on an Intel Core-i5 CPU with single core. We will make comparisons on execution time excluding deep learning methods as they all use GPUs for computation, which will make no sense for comparison. The execution time of several methods on Caltech-USA dataset is shown below (run time may vary on other different machines).

It is obviously illustrated in Figure 17 that ACF has much less run time while achieves much higher MR. The red dotted line is a dividing line, the lower right corner is, the performance the better. Spatial-Pooling and Checkerboards filters method have much larger run time (eight times slower than our method at least) while it may prevail in detection performance. These two methods both do not achieve a trade-off between performance and running time cost. Our method has nearly same execution time cost with LDCF because features' size in our method is identical to that of LDCF, while we achieve improved performance than LDCF. Therefore, our method accomplishes optimal balance between detection accuracy and execution time among all the non-deep state of the art pedestrian detection methods.

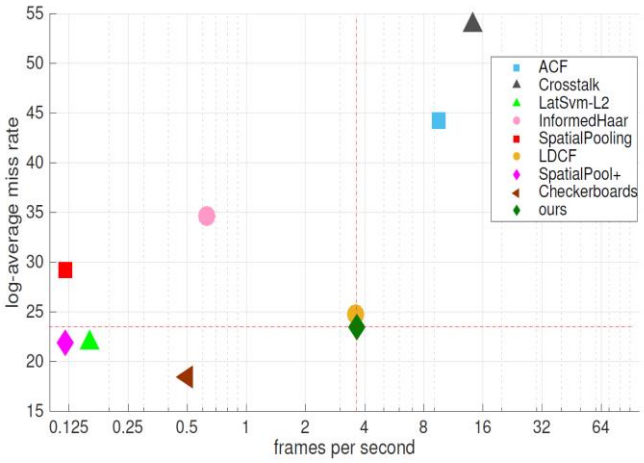



Figure 17. Running time versus detection performance.

Some of the sample detections have been show in Figure 18. Owning to concentration on feature extraction of head region, our method has an improvement on robustness to occlusion and deformation. The false positive detection could also be eliminated employing boost method as [10].

a) Detection on INRIA dataset.

b) Detection on Caltech-USA dataset.

Figure 18. Detection result samples.

## 5. Conclusions

By combining the traditional decorrelation operation and the feature extraction method based on the prior information of statistical human shape, according to the weight graph of different region features in the decision tree in the classification stage, the region with the highest response in the average human shape template, namely the head shoulder region, is selected to extract the covariance matrix, and the generated filter is shared within grouping channel. Based on the fact that the detection ability of local area learning filter is not directly proportional to its area size, the covariance matrix extracted from a specific area can contain more high identification pedestrian information, so as to improve the performance of the detector. Extensive experiments on INRIA and Caltech USA public pedestrian data sets show that the head shoulder region extraction filter can capture the deeper expression features of pedestrians, ensure the integrity of pedestrian information, and achieve the best balance between detection speed and detection accuracy.

## References


[1] Al-Najdawi N., Tedmori S., Edirisinghe E., and Bez H., "An Automated Real-Time People Tracking System Based on KLT Features Detection," *The International Arab Journal of Information Technology*, vol. 9, no. 1, pp. 100-107, 2012.

[2] Baek J., Kim J., and Kim E., "Fast and Efficient Pedestrian Detection via the Cascade Implementation of an Additive Kernel Support Vector Machine," *Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 902-916, 2017.

[3] Bastian B. and Jiji C., "Aggregated Channel Features with Optimum Parameters for Pedestrian Detection," *in Proceedings of International Conference on Pattern Recognition and Machine Intelligence*, Kolkata, pp. 155-161, 2017.

[4] Cao J., Pang Y., and Li X., "Pedestrian Detection Inspired By Appearance Constancy and Shape Symmetry," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5538-5551, 2016.

[5] Dollár P., Tu Z., Perona P., and Belongie S., "Integral Channel Features," *in Proceedings of British Machine Vision Conference*, London, pp. 91-98, 2009.

[6] Dollár P., Wojek C., Schiele B., and Perona P., "Pedestrian Detection: An Evaluation of the State of the Art," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743-761, 2012.

[7] Du X., El-Khamy M., Lee J., and Davis L., "Fused Dnn: A Deep Neural Network Fusion Approach to Fast and Robust Pedestrian Detection," *in Proceedings of IEEE Winter Conference on Applications of Computer Vision*, Santa Rosa, pp. 953-961, 2017.

[8] Lim J., Zitnick C., and Dollár P., "Sketch Tokens: A Learned Mid-Level Representation for Contour and Object Detection," *in Proceedings/CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, pp. 3158-3165, 2013.

[9] Nam W., Dollár P., and Han J., "Local Decorrelation for Improved Detection," *in Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montréal, pp. 424-432, 2014.

[10] Ohn-Bar E. and Trivedi M., "To Boost or Not to Boost? on The Limits of Boosted Trees for Object Detection," *in Proceedings of International Conference on Pattern Recognition*, Cancun, pp. 3350-3355, 2016.

[11] Ouyang W., Zhou H., Li H., Li Q., Yan J., and Wang X., "Jointly Learning Deep Features, Deformable Parts, Occlusion and Classification for Pedestrian Detection," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1874-1887, 2017.

[12] Shen J., Zuo X., Li J., Yang W., and Ling H., "A Novel Pixel Neighborhood Differential Statistic Feature for Pedestrian and Face Detection," *Pattern Recognition*, vol. 63, pp. 127-138, 2017.

[13] Shen J., Zuo X., Yang W., Yu H., and Liu G., "Learning Discriminative Shape Statistics Distribution Features for Pedestrian Detection," *Neurocomputing*, vol. 184, pp. 66-77, 2016.

[14] Tian Z., Shen C., Chen H., and He T., "FCOS: Fully Convolutional One-Stage Object Detection," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, pp. 9627-9636, 2019.

[15] Viola P. and Jones M., "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp.137-154, 2004.

[16] Zhang L., Lin l., Liang X., and He K., "Is Faster R-Cnn Doing Well for Pedestrian Detection?," *in Proceedings of European Conference on Computer Vision*, Amsterdam, pp. 443-457, 2016.

[17] Zhang S., Bauckhage C., and Cremers A., "Efficient Pedestrian Detection via Rectangular Features Based on a Statistical Shape Model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 763-775, 2014.

[18] Zhang S., Benenson R., and Schiele B., "Filtered Channel Features for Pedestrian Detection," *Computer Vision and Pattern Recognition*, vol. 1, no. 2, pp. 1751-1760, 2015.

[19] Zhang S., Benenson R., Omran M., Hosang J., and Schiele B., "How Far are we from Solving Pedestrian Detection?" *in Proceedings of Computer Vision and Pattern Recognition*, Las Vegas, pp. 1259-1267, 2016.

[20] Zhang S., Benenson R., Omran M., Hosang J., and Schiele B., "Towards Reaching Human Performance in Pedestrian Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 973-986, 2017.

[21] Zhao Y., Yuan Z., Chen D., Lyu J., and Liu T., "Fast Pedestrian Detection via Random Projection Features with Shape Prior," *in Proceedings of* IEEE *Winter Conference on Applications of Computer Vision*, Santa Rosa, pp. 962-970, 2017.

[22] Zhou X., Wang D., and Krähenbühl P., "Objects as Points," *in Proceedings of Computer Vision and Pattern Recognition*, Long Beach, pp. 141-183, 2019.

**Xue-ming Ding** received the Ph.D. Degree in control science and engineering from University of Science and Technology of China in 2005. He is currently an associate professor with University of Shanghai for Science and Technology. His research interests include system identification, embedded system and machine learning.

**Dong-fei Ji** received his first degree and M.S. degree both from the University of Shanghai for Science and Technology, Shanghai, China, all in control science and engineering. His interested topics are optical character recognition, pedestrian detection, pedestrian ReID and machine learning.