Color Maximal-Dissimilarity Pattern for Pedestrian Detection

Qingyuan Wang[1,2], Junbiao Pang[3], Guoyi Liu[4], Lei Qin[2], Qingming Huang[1,2], Shuqiang Jiang[2]
[1]*Graduate University of Chinese Academy of Sciences, China*
[2]*Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, China*
[3]*College of Computer Science and Technology, Beijing University of Technology, China*
[4]*NEC Labs China, Beijing, China*
[1,2]{*qywang, lqin, qmhuang, sqjiang*}*@jdl.ac.cn,* [3]*jbpang@bjut.edu.cn,* [4]*liu_guoyi@nec.cn*

Abstract

Feature plays an important role in pedestrian detection, and considerable progress has been made on shape-based descriptors. However, color cues have barely been devoted to detection tasks, seemingly due to the variable appearance of pedestrians. In this paper, Color Maximal-Dissimilarity Pattern (CMDP) is proposed to encode color cues by two core operations, i.e., oriented filtering and max-pooling, which emulate the functions of primary visual cortex (V1). The extensively experimental results reveal that the biologically-explainable encoding scheme increases the invariance of color cues, and outperforms the state-of-the-art color descriptor in terms of both accuracy and speed.

1. Introduction

Accurate pedestrian detection has immediate and far reaching impact to applications, such as intelligence surveillance and driver assistance systems. Yet detecting pedestrians in images is still a challenging task, because it is difficult to handle variable appearances and wide range of poses.

Currently gradient cues [1, 2] have been successfully used as shape information for pedestrian detection, but color cues barely attract enough attention. Only a few works [6, 9] have attempt to incorporate color cues. Schwartz et al. [6] propose a color descriptor called color frequency, which is captured by the number of times each color band is chosen, as the similar manners that HOG [1] choose the orientation of the gradient for a pixel. Walk et al. [9] introduce a descriptor, termed CSS, based on self-similarity of color values, i.e., similarities between colors in all different sub-regions, as they observe that colors of pedestrians are globally similar, e.g., color of face is similar to the one of hands.

Although these works [6, 9] indicate that color cues may help pedestrian detection, they do not answer the following important problems: 1) how to reduce the variable appearances, i.e., diverse colors of clothing; 2) how to handle the variations caused by poses. In this paper, we propose a color descriptor called Color Maximal-Dissimilarity Pattern (CMDP) to these problems based on biological knowledge.

In the experiments, the INRIA pedestrian dataset and PASCAL VOC 2007 dataset are served as our primary testbed. The experimental results on color cues reveal that CMDP achieves promising results in terms of both accuracy and speed. Because CMDP can capture some structure information via localized color distributions, thus it is independent of the pixel-level color. Moreover, CMDP is constructed on a hierarchical framework, and the two core operations, i.e., oriented filtering and max-pooling, make it invariant to shift and pose variations. In addition, combining CMDP with shape-based descriptor [1] improves the detecting performance. The promising results show that color cues are complement with shape cues and color cues can play an important role in pedestrian detection.

2. CMDP

In this section, CMDP is inspired by some biological knowledge to handle the variable appearances and pose variations of pedestrians.

2.1. Analogues between V1 and CMDP

Humans or primates outperform the best machine vision systems with respect to almost any measures, so building features that emulate object recognition in cortex has been an attractive goal. In the primary visual cortex (V1), the simple cells respond selectively to lines

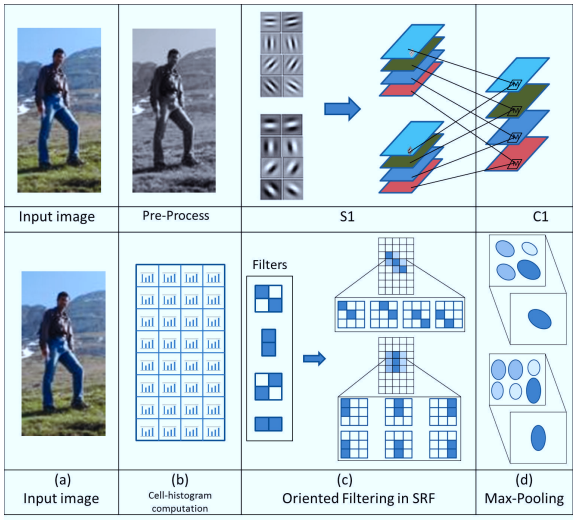


Figure 1: The analogues between V1 and CMDP. We use oriented gabor to simulate functions in V1. In (c), dissimilarity values are calculated by each filter in SRF, and we show "135°" and "90°" filters as example. In (d), we indicate the dissimilarity value with different size and color in eclipses, and the maximal value is selected over CRF.

or edges at particular orientations in a small receptive fields, termed as Simple Receptive Field (SRF) in this paper. While, the complex cells tend to have larger receptive fields (twice as large as simple cells, and termed as Complex Receptive Field (CRF) in this paper), respond to oriented bars or edges anywhere within their CRFs. Therefore, the function of complex cells is invariant to shift. In addition, SRF could be combined to form CRF [1].

In essence, there are two key procedures to simulate the orientation-selective feature detectors: 1) feature filtering; 2) max-pooling.

1. The feature filtering describes the basis of image wold, for instance, Gabor filters (See S1 in Figure 1) can well simulate these oriented response in visual cortex, and thus are widely used in computer vision [4].

2. The max-pooling (See C1 in Figure 1) not only reduces the amounts of information passed to next visual stage, but also increase the invariant ability of features [3].

As illustrated in Figure 1, 4 filters are designed in CMDP, and two of them are selected as example to illustrate the usage and pooling of filters in (c) and (d), respectively. Thus, CMDP tries to simulate the core elements of orientation-selective feature detectors in V1. So, CMDP is a particular V1-like feature.

[1]For more details, please read reference [8].

2.2. Details of CMDP

Now we give implementation details of CMDP and introduce CMDP in the extraction order as Figure 1 showed.

Cell-histogram Computation. This procedure is implemented by dividing the image window into small spatially non-overlapped regions ("cell"), showed in Figure 1(b). For each cell a local color histogram is accumulated over every color channel. To reduce quantization effects, trilinear interpolation is applied. Finally, the histogram representation of each cell is achieved by concatenating the histograms of different color channels and then cell-normalization is carried out. So, the output of this procedure is a cell-histogram feature map.

There are some parameters affecting the descriptor's performance, including the color space and the histogram dimension. Following the conclusions in [5], CIE-LUV color space and 10-dimensional histogram per channel are chosen in CMDP.

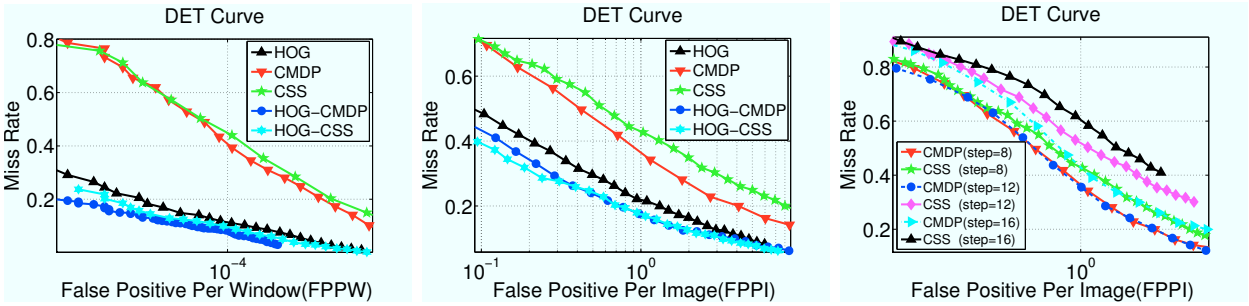
Oriented Filtering in SRF. In this procedure, we define four orientated filters (see Figure 1(c)). In our framework, as an analogue of simple cells, the filters are designed to capture edge information in the simple receptive field. And, the size of SRF is determined by the size of filters.

In each filter, the dissimilarity value is calculated between two adjacent cells (indicated by the same color in Figure 1(c)), as:

$$d(i,j,k) = \text{dissim}(cell_{i,j}, cell_{i_k,j_k}), k \in \{1,2,3,4\}$$
$$\text{s.t.} \begin{cases} i_1 = i-1, j_1 = j-1; i_2 = i-1, j_2 = j; \\ i_3 = i-1, j_3 = j+1; i_4 = i, j_4 = j+1. \end{cases} \quad (1)$$

where $cell_{i,j}$ represents an n-dimension color histogram for cell in row i and column j. k indicates different filters showed in Figure 1(c). (i_k, j_k) is the adjacent cell's offset location relative to $cell_{i,j}$ for different filters. And dissim$(\cdot)$ is a function for measuring the dissimilarity between two different histograms. We experimented with a number of well-known distance functions including the L1-norm, L2-norm, χ^2-distance, and histogram intersection to calculate the dissimilarity values. Finally, we choose histogram intersection as the comparison results suggest.

Max-pooling in CRF. Physiological studies have shown that suppression effect exists between neurons in CRF[8]. So in our implementation the corresponding pooling operation is a Max operation. The max pooling can increase spatial-invariance of feature.



(a) (b) (c)

Figure 2: Comparisons between CSS and CMDP.

We pool the maximal dissimilarity value in the same orientation filtering the CRF showed in Figure 1(d). The max-pooling operation is:

$$H(i,j,k) = \max_{m,n}(d(m,n,k)), k \in \{1,2,3,4\}$$
$$\text{s.t.} \begin{cases} if\ \ k=1, m \in \{i, i+1\}, n \in \{j, j+1\}; \\ if\ \ k=2, m \in \{i, i+1\}, n \in \{j-1, j, j+1\}; \\ if\ \ k=3, m \in \{i, i+1\}, n \in \{j-1, j\}; \\ if\ \ k=4, m \in \{i-1, i, i+1\}, n \in \{j, j+1\}. \end{cases} \quad (2)$$

Local Normalization. The max-pooled histograms are then normalized in local neighboring cells. Firstly, the "energy" of each cell is calculated as following [10]:

$$E(i,j) = \sum_{k=1}^{4} H(i,j,k)^2 \quad (3)$$

Secondly, we can get the "energy" of the local area by summing all the "energy" of the cells in this area. Then, the normalized histogram is computed by

$$N(i,j,k) = \frac{H(i,j,k)}{\sum_{m=i-1}^{i+1}\sum_{n=j-1}^{j+1} E(m,n)} \quad (4)$$

The final histogram is truncated by value σ for better performance. For CMDP, we find that σ=0.2 is a good choice.

3. Experiments

In our experiments, we will evaluate CMDP on INRIA pedestrian dataset and the challenging PASCAL VOC 2007 dataset. We utilize the per-window and per-image evaluation methodologies for the INRIA dataset, while average precision criterion is adopted for the PASCAL VOC dataset experiment.

For INRIA dataset, we use the stochastic optimization based linear SVM, i.e., Pegasos [7], to train the detectors, due to the large number of training examples. During the training procedure, one round of retraining (bootstrapping) protocol is utilized as Dalal et al. suggested in [1]. While for PASCAL VOC dataset, part based model using latent SVM as Felzenszwalb described in [2] is applied.

3.1. Experiments on INRIA Dataset

In this subsection, we will study the performance of CMDP on INRIA dataset, and draw comparisons with CSS [9], which is the current state-of-the-art color descriptor on pedestrian detection. Some comparison results are showed in Figure 2.

Figure 2(a) and 2(b) are the per-window and per-image results respectively, where the sliding step is 8 pixels and the scale ratio is 1.2. From Figure 2(a) and 2(b), we can observe that CMDP outperforms CSS. Moreover, the performance of CMDP is still comparative with CSS when combined with HOG[1].

Table 1: Detailed comparisons between CSS and CMDP (The value for different steps is the miss rate when FPPI = 1).

Features	dimensions	step=8	step=12	step=16
CMDP	512	35.8%	35.7%	42.8%
CSS	8128	42.9%	51.8%	60.4%

The comparison result is amazing, because CMDP is only a 512-dimensional descriptor, while CSS is 8128-dimensional for a 128×64 window. Moreover, the complexity of CMDP is lower than CSS, because the times of the similarity calculation of CMDP is far fewer than CSS. In addition, in order to specify the translation invariance of CMDP, we visually show the comparison result under different sliding steps in Figure 2(c). The detailed comparisons between CMDP and CSS is showed in table1. So, when we change the sliding step, the performance of CMDP declines much slowly than CSS. Specially, the performance of CMDP is almost the same when step=8 and step=12.

In addition, the average feature maps of INRIA dataset corresponding to the 4 orientations are showed

Table 2: Evaluation results on PASCAL VOC 2007 Dataset.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	MAP
V4[2]	28.9	59.5	10	15.2	25.5	49.6	57.9	19.3	22.4	**25.2**	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.28
Proposed	**30.6**	**60.2**	**11**	**16.8**	**25.9**	**50.5**	**58.3**	**20.8**	**23.5**	25.1	**24.4**	**11.7**	**58**	48.7	**42.2**	**13.7**	**19.9**	**34.2**	**46.7**	**43.2**	**33.27**

in Figure 3. This figure clearly reveals the shift invariance property of CMDP. Because of the max-pooling technique in the encoding scheme, CMDP can capture the structure information around pedestrians coarsely to increase the shift invariance.

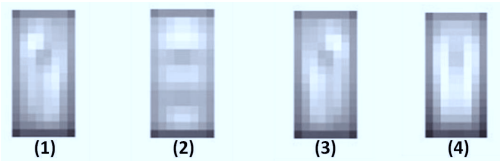


Figure 3: The positive mean feature maps of INRIA dataset

In pedestrian detection, HOG is a pixel-level feature, and it is designed to capture subtle shape information locally. So, the shape description by HOG is easily affected by "noises" and the robustness of HOG in crowed environment is relatively low. While CMDP is a patch-level feature and is less sensitive to "noises". Consequently, CMDP can complement with HOG well. We show some false negatives generated by HOG, but successfully detected by HOG-CMDP in Figure 4.


Figure 4: Some false negatives eliminated by CMDP.

3.2. Experiments on PASCAL VOC Dataset

In this part of experiment, we choose latent SVM as our detector. Several versions of latent SVM were released at Felzenszwalb's homepage and we use the latest one called VOC-release4 [2] as V4 shortly. In addition, we use the following scheme to add CMDP to the framework of V4: using the single feature HOG to train root and part, and after we have got the aligned samples, CMDP is added to the model to retrain. The comparison results are showed in Table 2.

From Table 2, we can observe that CMDP can improve the detection performance on almost all categories except "cow" and "motorbike". The result is so promising because CMDP has only 4 dimension in each cell; while, HOG used in V4 is 32 dimensions. Therefore, we can use CMDP to achieve satisfactory improvement with little cost.

4. Conclusion

In this paper, we propose a novel biologically-explainable color descriptor, CMDP, for pedestrian detection. The experimental results show that the orientation filtering in CMDP can effectively capture the local structure information, and max-pooling can increase the invariance of shift. Therefore, CMDP is an useful color descriptor and color cues indeed play an important role in pedestrian detection.

Acknowledgements

This work was supported in part by National Basic Research Program of China (973 Program): 2009CB320906, in part by National Natural Science Foundation of China: 61025011, 61035001, 61133003 and 61003165, and in part by Beijing Natural Science Foundation: 4111003.

References

[1] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. *CVPR*, 2005.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010.

[3] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *in The Handbook of Brain Theory and Neural Networks, MA: MIT Press*, 1995.

[4] Neumann, H. N. Sepp, and W. Sepp. Recurrent v1-v2 interaction in early visual boundary processing. *Biological Cybernetics*, 1999.

[5] Q.Wang, J.Pang, L.Qin, S.Jiang, and Q.Huang. Justifying the importance of color cues in object detection: a case study on pedestrian. *PCM*, 2011.

[6] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. *ICCV*, 2009.

[7] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. *ICML*, 2007.

[8] T.Serre, L. Wolf, and T. Poggio. Robust object recognition with cortex-like mechanisms. *TPAMI*, 2007.

[9] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. *CVPR*, 2010.

[10] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structured hog-lbp for object localization. *CVPR*, 2011.