Justifying the importance of color cues in object detection: a case study on pedestrian

Qingyuan Wang[1,2], Junbiao Pang[3], Lei Qin[2],
Shuqiang Jiang[2], Qingming Huang[1, 2]

[1]Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
[2]Key Lab of Intell.Info.Process., Inst. of Comput. Tech., CAS, Beijing 100190, China
[3]Beijing Municipal Key Lab. of Multimedia and Intelligent Software Technology, College of Computer Science and Technology, Beijing University of Technology, 100124, China
{qywang, jbpang, lqin, sqjiang, qmhuang}@jdl.ac.cn

Abstract. Considerable progress has been made on hand-crafted features in object detection, while little effort has been devoted to make use of the color cues. In this paper, we study the role of color cues in detection via a representative object, i.e., pedestrian, as its variability of pose or appearance is very common for "general" objects. The efficiency of color space is first ranked by empirical comparisons among typical ones. Furthermore, a color descriptor, called MDST (Max DisSimilarity of different Templates), is built on those selected color spaces to explore invariant ability and discriminative power of color cues. The extensive experiments reveal two facts: one is that the choice of color spaces has a great influence on performance; another is that MDST achieves better results than the state-of-the-art color feature for pedestrian detection in terms of both accuracy and speed.

Keywords: Pedestrian detection, Color space analysis, Color descriptor, Invariant descriptor.

1 Introduction

In recent years, detecting the predefined objects in images has been an attracting problem in computer vision, e.g., face [13, 14, 15], pedestrian [4, 5, 6, 7, 8, 9]. Accurate detection would have immediate impacts on many real applications, such as intelligence surveillance [1] and driver assistance systems [2, 3]. Detecting non-rigid objects in images is still a challenging task, due to the variable appearance and the wide range of poses. In this paper, pedestrian is used as a typical case, because its variable poses and diversity of clothing are representative and challenging for "general" objects, e.g., animals, car, et [16].

Recently shape information [4, 5, 6], mainly described by gradients, has been exhaustively explored and successfully used in detection, and yet color cues do not attract

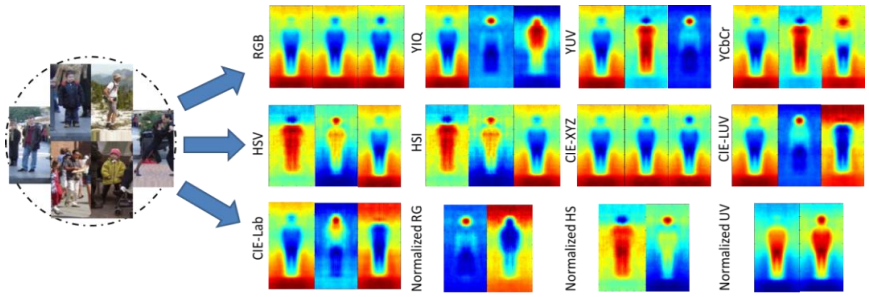

Fig. 1. The average positive examples of INRIA dataset are visualized in the different color spaces and the corresponding channels. Some images show clear silhouette-like boundary, while other just enhances parts of the pedestrians, e.g., head in YIQ space.

enough attention. Moreover, color cues have been facilely considered to be useless, due to the variability in the color of clothing. In our best knowledge, we first try our best to systematically evaluate the role of color cues, and show that color cues can achieve competitive results in pedestrian detection. Our study is based on the following methodology: 1) There are a variety of color spaces, yet which one is better for detection is still a controversy problem [8, 9]; therefore, typical color spaces are naturally compared (see Fig. 1.). 2) A simple yet efficient color descriptor is proposed to be evaluated on the selected color spaces.

Although the INRIA pedestrian dataset [4], as our primary testbed, may be relatively simple [9], our experiments about color cues still reap a huge harvest. First, we show that there are significant differences of performance among color spaces, thus, the choice of color spaces is very important in detection. Second, different coding schemes lead to diverse accuracies, and thus the selection of coding method is a key to boost performance. Third, the combination of color descriptors with the shape-based feature [4] shows the complementary yet important role of color cues in object detection.

The remainder of this paper is organized as follows. In Sec. 2, we give an overview of the related works. In Sec. 3, we introduce different color spaces and describe the details on evaluating color spaces. Then our proposed color descriptor will be presented in Sec. 4. In Sec. 5 we perform the detailed experimental evaluations. Finally, we conclude this paper in section 6.

2 Related Work

Color information is very popular in image classification [10][22], but it does not attract enough attention in object detection, and most people still doubt the efficiency of color for detection. On the other side, a few literatures use color cues in naïve approaches, and set a minor role for color in detection.

In HOG [4], the orientation of the gradient for a pixel is selected from the R, G and B channels according to the highest gradient magnitude. Thus some color information can be captured by the number of a channel is chosen in HOG [4]. Comparing with HOG, Schwartz et al. [7], further, construct a three bin histogram that tabulates the number of times each color channel is chosen. So they call this color descriptor as color frequency. This color feature mainly captures the color information of faces, due to the possible homologous skin color in faces. However, the information of whole body tends to be ignored as illustrated in Fig. 1.

Walk et al. [8] observe that the colors of pedestrians may be globally similar, e.g., the color of faces is similar to the one of hands. Therefore, color self-similarity (CSS) is introduced to capture the similarities among whole body. This feature captures pair-wise spatially statistics of color distribution, e.g., color of clothes ignored by the color frequency descriptor [7]. On the other hand, the computational cost of CSS is intensive, because CSS calculates the global self-similarity within the detection window. For instance, the dimension of CSS for a 128x64 window is 8,128.

Both of the above works are integrated with Support Vector Machine (SVM) classification framework, while Dollar et al. [9] use color cues in boosting via an integral channel features approach. The color features are firstly generated by summing the pixel values in the pre-designed templates. The discriminative feature is chosen by boosting classifiers. It is a type of first-order feature, and the disadvantages of the color frequency descriptor [7] still exist in this coding method. On the other interesting side, Dollar et al. [9] and Walk et al. [8] give totally different conclusion on using color spaces for object detection. In this paper, the efficiency of color spaces will be first systematically ranked by empirical comparisons.

3 Analysis of Color Space

In this section, the typical color spaces are introduced, and then evaluated to rank the performance of color spaces for pedestrian detection.

3.1 Color Spaces

Color is usually described by the combined effect of three independent attributes /channels. For example, R, G and B are the widely used color channels. There are many other color spaces in computer vision or related communities, because different color spaces can describe the same object from different perspectives. In this paper, we divide the color spaces into three categories according to the application purposes.

Color spaces for Computer Graphics. These color spaces are mainly used in display system, for example, RGB, HSV and HSI. And RGB, HSV and HSI are chosen to be evaluated in our experiment.

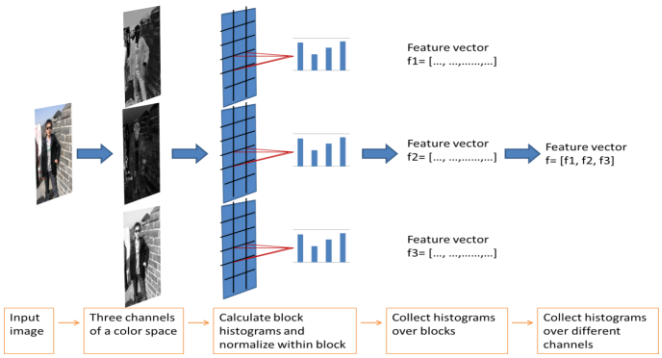

Fig. 2. The HLC extract procedure. A three-channel color space is used as an example.

Color Spaces for TV System. YIQ and YUV are analogue spaces for NTSC (National Television System Committee) [17] and PAL (Phase Alternating Line) [18] systems respectively, while YCbCr is a digital standard. And YIQ, YUV and YCbCr are all picked in our experiment.

CIE Color Spaces. The CIE (International Commission on Illumination) [19] system characterizes colors by a luminance parameter Y, and two color coordinates x and y. Some typical CIE spaces, i.e., XYZ, LUV and Lab, will be evaluated.

In addition, in many applications in order to keep the luminance invariance, the intensity attribute is discarded and only the chrominance is kept. So, in our experiment we will also evaluate this kind of color spaces, including normalized RG, normalized HS and normalized UV [8].

3.2 Evaluation of color spaces

One of our goals is to find and explain that which color space would be the best one for pedestrian detection. On the other side, the color feature may have great influence on the accuracy, and thus feature should be simple and unbiased to reflect the ability of color spaces. As illustrated in Fig. 1, color histogram may be a good choice to describe the distribution of colors. A localized color histograms, termed HLC (Histogram of Local Color), is advocated as our evaluation feature.

HLC. It firstly converts an image into different channels, and then divides each color channel image into small spatial non-overlapped regions ("block"). In each block, a local 1-D histogram of color values is accumulated[1]. Next, the channel-level HLC is constructed by concatenating the histograms in all blocks. The image-level HLC is finally obtained by concatenating the channel-level HLCs (see Fig. 2). In our experiments, we compute histograms on 16×16 pixels blocks.

[1]During histogram calculation, trilinear interpolation is applied to avoid quantization effects.

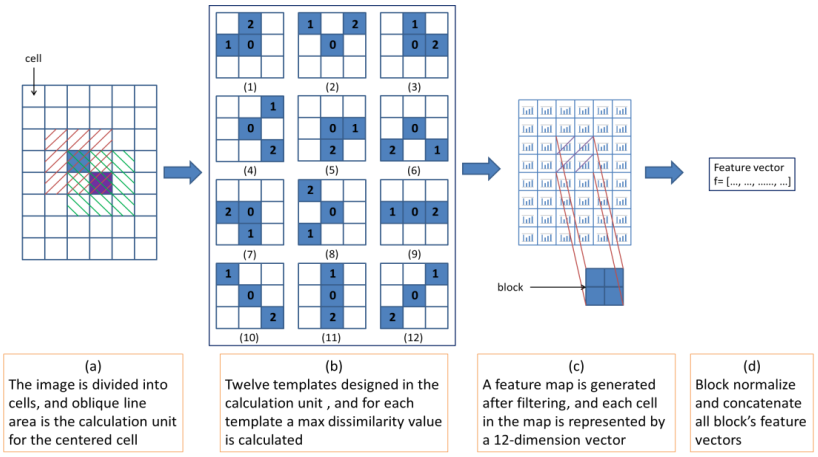

Fig. 3. The method of calculating the MDST for an example

4 Max DisSimilarity of Different Templates

As shown in Fig. 1, the boundary of the whole or parts of pedestrian can be clearly illustrated. Our work is to explore a descriptor that can capture the boundary effectively regardless of variability of clothing or its color. Max DisSimilarity of different Templates (MDST), based on the max-coding, is designed to capture the local boundary and to increase the translation invariance of feature.

MDST. Firstly, the image is divided into non-overlapped regions (we call it as "cells"). The feature/"attribute" of a cell is calculated in the enlarged local area/"calculation unit"[2] which centers on the cell, as shown in Fig. 3(a). In the calculation unit, 12 templates, shown in Fig. 3(b), are designed to capture the local boundary information. The number 0, 1 and 2 indicate three different cells in a template. For the j-th cell in the i-th template, an n-dimension color histogram, $cell_{ij}$, is extracted to calculate the dissimilarity values.

$$ds_{ij} = dissim(cell_{i0}, cell_{ij}) \quad i = 1,2,3...,12; j = \{1,2\} \tag{1}$$

Where "$dissim$" is a function to measure the dissimilarity between two cells. There are many possibilities to define dissimilarity for histogram comparison. In our experiments, a number of well-known distance functions, L1-norm, L2-norm, chi-square-distance, and histogram intersection, are all evaluated, and histogram intersection works best.

The dissimilarity scores, mds_i , in the i-th template is calculated by max-pooling between 0-1and 0-2 dissimilarity pairs,

[2]In our experiment, we use a 3x3-cell to be the calculation unit as the oblique line indicates.

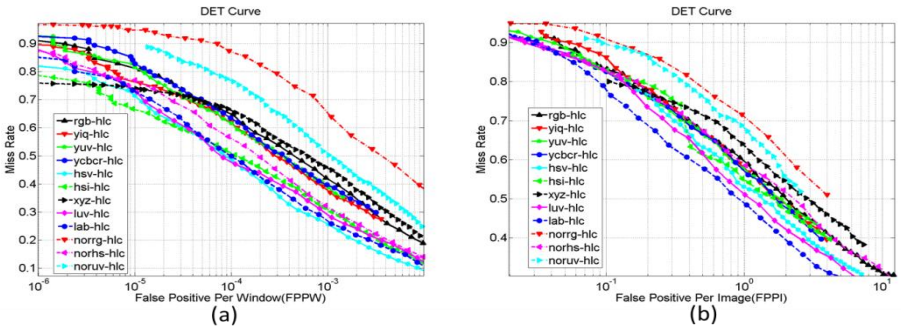

Fig. 4. Evaluation of diverse color spaces on HLC. (a) The per-window evaluation result. (b) The per-image evaluation result.

$$mds_i = \max(ds_{i1}, ds_{i2}) \quad i = 1, 2, 3..., 12 \tag{2}$$

So, the feature for a cell is extracted by assembling the score mds_i in all 12 templates as

$$f = [mds_1, mds_2, mds_3, ..., mds_{12}]. \tag{3}$$

The feature vectors further are normalized in a 2x2 cell (termed as "block"), as illustrated in Fig. 3(c). We get a 48-dimension feature over one block after the block-wise normalization, and the MDST can concatenate features in all blocks. In our experiments, the cell size is 8x8 pixels for a 128x64 window, which would have 1,536-dimension MDST.

5 Experimental Results and Discussions

In our experiment, we utilize two evaluation protocols that are the per-window and per-image evaluation. The per-window evaluation methodology is to classify manually normalized examples against windows sampled at a fixed density from images without pedestrians, corresponding to the miss rate against False Positive Per Window (FPPW). While in the per-image evaluation, a detector is densely scanned across an image with or without pedestrians and nearby detections merged using non maximal suppression (NMS), corresponding to the miss rate against False Positive Per Image (FPPI).

The per-window evaluation can avoid NMS, however, it does not measure errors caused by detecting at incorrect scales, positions or false detections arising from body parts [11]. While the per-image evaluation method takes into consideration of all the impact factors. In addition, our experiment comparisons are mainly under 10^{-4} FPPW and 10^{0} FPPI, which is adopted by most evaluation schemes in pedestrian detection.

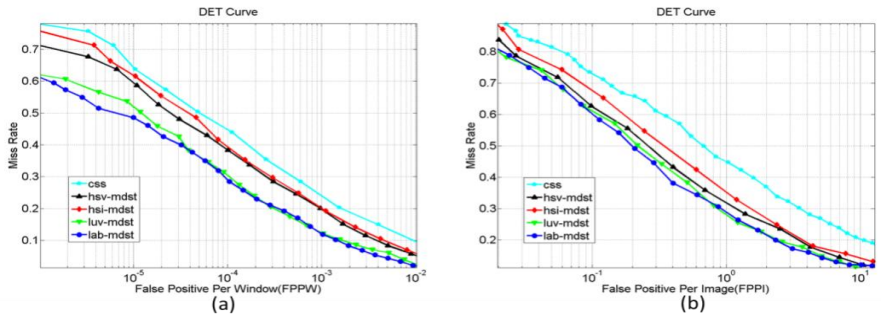

Fig. 5. Comparisons between MDST and CSS. (a) The per-window evaluation result. (b) The per-image evaluation result.

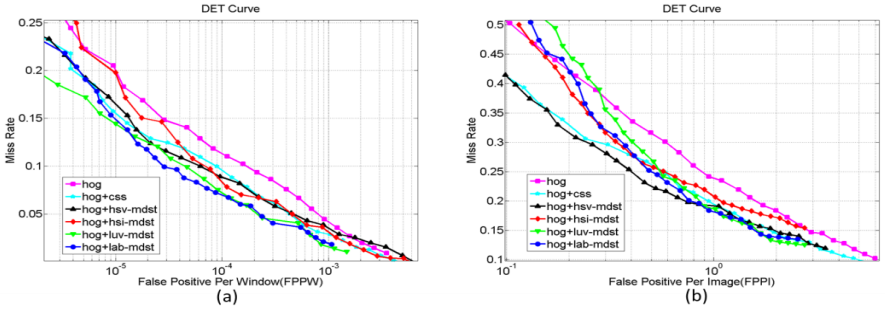

Fig. 6. Comparisons between MDST+HOG and CSS+HOG. (a) The per-window evaluation result. (b) The per-image evaluation result.

For all subsequent experiments, we use the stochastic optimization based linear SVM [12], i.e., Pegasos, 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 [4].

5.1 Evaluation of color space

In Figs. 4(a) and 4(b), we plot the performance of HLC on various color spaces according to the per-window and per-image evaluations. This figure clearly shows that the perspective of describing colors has great influences on the performance.

Firstly, the performance of normalized RG, normalized HS and normalized UV [8] are all worse than the corresponding RGB, HSV and YUV. According to the fact that the human eyes have three different types of color sensitive cones [20], the response of the eyes is best described in terms of three "tristimulus values". Therefore, the intensity attribute of a color space, ignored by normalized color space, may be critical to describe pedestrian.

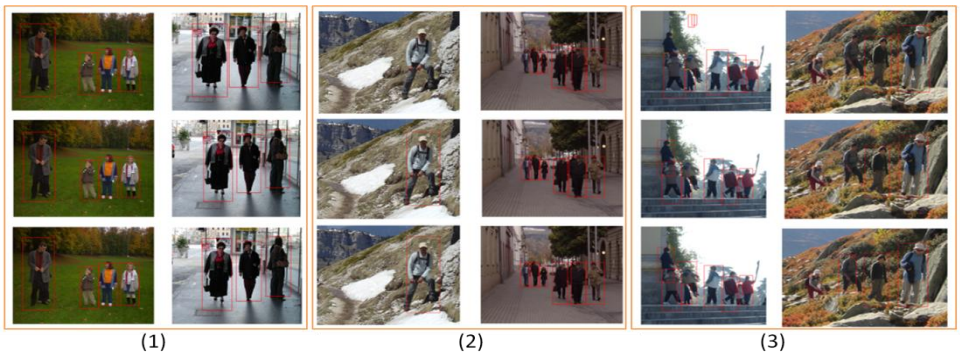

Fig. 7. The first row is the detection results of HOG. The second row is the detection results of HOG-CSS (HSV) and the third row corresponds to HOG-MDST (CIE-Lab) at FPPI = 10^0.

Secondly, the four best performed spaces are ranked in a descendent order, CIE-Lab, CIE-LUV, HSV and HSI. The color spaces in TV system are not as discriminative as any one of the four ones. On the contrary, CIE-based color spaces (CIE-Lab and CIE-LUV) are designed to approximate human vision [19]; HSV and HSI are relevant to the perception of human eyes [21]. Therefore, the color spaces，CIE-Lab, CIE-LUV, HSV and HSI, associate with the human perception, and achieve better performance for detection.

5.2 The performance of MDST

To consider the influence of the color spaces on descriptors, we plot the per-window and per-image evaluation curves of the MDST in Fig. 5. Not surprisingly, the performances of the MDST on four color spaces are consistent with the results of our color space evaluation experiment. Therefore, CIE-Lab/CIE-LUV color spaces could be the best ones for pedestrian detection.

In these two Figs. 5(a) and 5(b), the performances of MDST are better than CSS (HSV as suggested in [8]), which is the current state-of-the-art color descriptor. In terms of computational cost, the dimension of MDST is 1,536, while the dimension of CSS is 8,128 for a 128x64 window. Thus, the MDST is a better color descriptor for its lower computational cost and better performance.

We will evaluate the augmented HOG-MDST feature to study whether the two descriptors can complement with each other or not, because Histogram of Oriented Gradients (HOG) [4] is the most successful shape descriptor in pedestrian detection. In Figs. 6(a) and 6(b), it can be seen that MDST raises detection rate 4% at 10^{-4} FPPW and 6% at 10^0 FPPI, comparing with HOG. On the other side, the performances of different color spaces are also consistent with the results of color space evaluation under 10^{-4} FPPW and 10^0 FPPI.

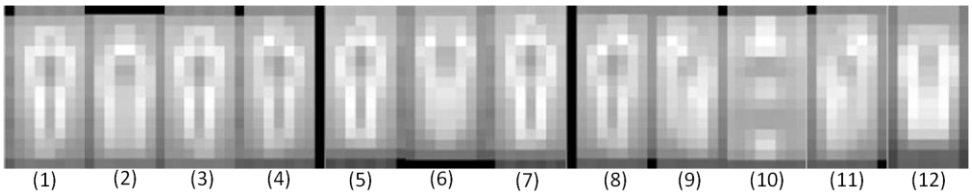

Fig. 8. The average feature map of the 12 templates for INRIA dataset

In the comparison between MDST and CSS, the MDST combined with HOG has better performance than CSS. In our experiment, the HOG-MDST (CIE-Lab) can achieve 86% detection rate at 10^0 FPPI; while the HOG-CSS (HSV) feature can only achieve 83.5% detection, raising 4.5%. For a more intuitive comparison, we show the detection results of HOG, HOG-CSS and HOG-MDST (CIE-Lab) in Fig. 7. Fig. 7(1) indicates that MDST has no negative effect on detection; while from Figs. 7(2) and 7(3), we can get that MDST has a superior accuracy than CSS in complex scenes.

5.3 Further discussions on MDST

The average feature maps of INRIA dataset are illustrated in Fig. 8. These average feature maps (1) ~ (12) correspond to the result of templates (1) ~ (12) in Fig. 3(c), respectively. These figures capture the coarse boundary around pedestrians. The color cues achieve translation invariance locally, because of the max-pooling used in these templates.

In addition, from Figs. 5(a), 5(b), 6(a) and 6(b), we find an interesting phenomenon, that is, the MDST(CIE-Lab) on its own achieves 18% higher accuracy than CSS(HSV) under FPPI; while, the difference between HOG-MDST(CIE-Lab) and HOG-CSS(HSV) is just around 2% at FPPI=1. There may be some redundancies between HOG and MDST, because MDST uses color cues to capture pedestrian boundary. However, MDST derived from color cues still provides complementary information for shape-based features.

6 Conclusions

In this paper, our experiment justifies the role of color in object detection, and shows color cues play an important role. First, the typical color spaces are systematically evaluated for pedestrian detection. The results of our evaluation experiments indicate that the choice of color spaces has a great influence on performance. CIE-Lab/CIE-LUV spaces, for pedestrian, are the most suitable color spaces. Secondly, the performance of the MDST shows that max-coding scheme can effectively capture the local boundary information and outperform the state-of-the-art.

In future work, we plan to evaluate MDST on more different objects, e.g., car, cat, house [16], and explain more insights why max-coding scheme is efficient for color cues.

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 and 61003165, and in part by Beijing Natural Science Foundation: 4111003.

References

1. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: 9th IEEE International Conference on Computer Vision, pp. 734--741(2003)
2. Wojek, C., Walk, S., Schiele, B.: Multi-Cue Onboard Pedestrian Detection. In: 23rd IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp.794--801(2010)
3. Geronimo, D., Lopez, A., Sappa, A.: Survey of Pedestrian Detection for Advanced Driver Assistance Systems. In: IEEE Transactions on Pattern Analysis and Machine Intelligence(2010)
4. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: 18th IEEE Conf. Computer Vision and Pattern Recognition, vol.1, pp. 886--893(2005)
5. Maji, S., Berg, A. C., Malik, J.: Classification Using Intersection Kernel Support Vector Machines is efficient. In: 21th IEEE Conf. Computer Vision and Pattern Recognition(2008)
6. Felzenszwalb, P., McAllester, D., Ramanan, D.: A Discriminatively Trained, Multi-scale, Deformable Part Model. In: 21th IEEE Conf. Computer Vision and Pattern Recognition(2008)
7. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.: Human detection using partial least squares analysis. In12th IEEE International Conference on Computer Vision, pp. 24--31(2009)
8. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and Insights for Pedestrian detection. In: 23rd IEEE Conf. Computer Vision and Pattern Recognition, pp.1030--1037(2010)
9. Dollar, P., Tu, Z., Pernoa, P., Belongie, S.: Integral channel features. In: 20th British Machine Vision Conference(2009)
10. Van de Sande, K. E. A., Gevers, T., Snoek, C. G. M.: Evaluation of color descriptors for object and scene recognition. In: 21th IEEE Conf. Computer Vision and Pattern Recognition(2008)
11. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: A Benchmark. In22nd IEEE Conf. Computer Vision and Pattern Recognition (2009)
12. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal Estimated sub-GrAdientSOlver for SVM. In: International conference on Machine learning (2007)
13. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (1998)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascadeof simple features. In14th IEEE Conf. Computer Vision and PatternRecognition, volume 1, pp. 511--518 (2001)
15. Huang, C., Ai, H., Li, Y., Lao, S.: Vector boosting for rotation invariantmulti-view face detection. In: 10th IEEE International Conference on Computer Vision, pages 446--453 (2005)
16. Everingham, M., Van~Gool, L., Williams, C. K. I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. In: International Journal of Computer Vision, jun (2010)
17. Wikipedia for NTSC information, http://en.wikipedia.org/wiki/NTSC
18. Wikipedia for PAL information, http://en.wikipedia.org/wiki/PAL
19. Wikipedia for CIE, http://en.wikipedia.org/wiki/International_Commission_on_Illumination
20. Ng, J., Bharach, A., Zhaoping, L.: A survey of architecture and function of the primary visual cortex. In: Eurasip Journal on Advances in Signal Processing (2007)
21. Wikipedia for HSL and HSV information, http://en.wikipedia.org/wiki/HSL_and_HSV
22. Gevers, T., Smeulders, A.: Color Based Object Recognition. In :Pattern Recognition, Volume 32, Pages 453--464 (1997)