Information Sciences 405 (2017) 190–206

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Rotative maximal pattern: A local coloring descriptor for object classification and recognition

Junbiao Pang[a], Jing Huang[a], Lei Qin[c], Weigang Zhang[d], Laiyun Qing[b,*], Qingming Huang[b,c], Baocai Yin[a]

[a] Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing, University of Technology, No.100 Pingleyuan Road, Chaoyang District 100124, China

[b] School of Computer and Control Engineering, University of Chinese Academy of Sciences, No.19 Yuquan Road, Shijingshan District, Beijing 100049, China

[c] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, No.6 Kexueyuan South Road, Haidian District, Beijing 100190, China

[d] School of Computer Science and Technology, Harbin Institute of Technology at Weihai, No. 2 West Wenhua Road, Weihai 26209, China

ARTICLE INFO

Article history:
Received 29 June 2016
Revised 5 April 2017
Accepted 7 April 2017
Available online 8 April 2017

Keywords:
Object detection
Object recognition
Max pooling
Translation invariance
Self similarity
Photometric invariance

ABSTRACT

Inspired by the photometric invariance of color space, this paper proposes a simple yet powerful descriptor for object detection and recognition, called Rotative Maximal Pattern (RMP). The effectiveness of RMP comes from the two components: Rotatable Couple Templates (RCTs) with max pooling, and Normalized Histogram Intersection (NHI) with the theoretical guarantee. More concretely, RCTs are the combination of two templates to code the possible rotations. NHI serves as the similarity between two color histograms. We have conducted extensive experiments on INRIA pedestrian and Pascal VOC2007 data sets for object detection tasks; we also show that our approach leads to a promising performance on Caltech 101, Scene 15, UIUCsport and Stanford 40 action data sets.

1. Introduction

Recently there has been much interest to use color information in visual object recognition and object detection, *e.g.*, local invariant features [28], Color Name (CN) [15]. There is also several work to evaluate the invariance and the discrimination of color descriptors. Essentially, the success of the invariant photometric of color heavily depends on the appearance of an object itself – whether color is the important cue to describe the object or not. If features from color barely capture the dominant pattern, the color-based descriptors tend to achieve inferior performances. Therefore, a successful approach combines the color cues with other complementary ones, in order to obtain more discriminative ability [37].

There are two ways to combine the complementary cues with color descriptors:

- *At the feature level:* The typical approach [1] combines color cues with Scale-Invariant Feature Transform (SIFT) [37].
- *At the classifier level:* The other cues and color descriptors are directly fused by classifiers, *e.g.*, Multiple Kernel Learning (MKL) [32] or boosting [8].

[*] Corresponding author.
E-mail addresses: junbiao_pang@bjut.edu.cn, pang.junbiao@gmail.com (J. Pang), jing.huang@vipl.ict.ac.cn (J. Huang), lqin@jdl.ac.cn (L. Qin), wgzhang@hit.edu.cn (W. Zhang), lyqing@ucas.ac.cn (L. Qing), qmhuang@jdl.ac.cn (Q. Huang), ybc@bjut.edu.cn (B. Yin).

http://dx.doi.org/10.1016/j.ins.2017.04.011
0020-0255/© 2017 Published by Elsevier Inc.

Technically, concatenating descriptors by the first approach results in a high-dimensional feature vector, naturally increasing both the storage requirement and the computational cost; while the second heavily depends on the generalization ability of classifiers.

Therefore, it is natural to ask whether color cues can be combined with other complementary cues in a more compact scheme without depending on the generalization of classifiers. There are many potential benefits of this scheme: reduced storage requirements, reduced computational complexity, and improved classification performances. A more compact descriptor would simplify both the pattern representations and the subsequent classifiers.

Inspired by the promising results from the fusion of both the shape-based features and the color-based ones [15], we seek an *invariant* descriptor with a *good* generalization ability, based on two motivations. First, although an enormous volume of literature has been devoted to feature fusion, there is little attention about how to obtain compacted descriptors. Second, we want to avoid the disadvantages of the traditional approaches [32]: the computational burden of the enlarged feature vector, and the requirement of the generalization of classifiers. In summary, we desire a computationally simple method to fuse multiple cues into a compacted descriptor.

In this paper, rather than adopting a solo filter (which is a common coding scheme [7], but is difficult to obtain invariant ability as will be analyzed), we propose Rotatable Couple Templates (RCTs) to achieve the invariance ability. RCTs design multiple templates to capture the possible rotations; this is a significant departure from the traditional framework [22]. Like other patch-wise features [36], RCTs compute the dissimilarity by Normalized Histogram Intersection (NHI). Finally, the responses of RCTs are assembled into compact Rotative Maximal Pattern (RMP) with max pooling.

To the best of our knowledge, this paper is the first to incorporate the coding schemes from shape descriptors into a compacted coloring feature. The proposed method is computationally simple, yet powerful. Simply by selecting a set of maximal responses from RCTs, with no further parameter tuning, we find a coloring descriptor meets or exceeds the current state-of-the-arts on different tasks.

The rest of this paper is organized as follows: in Section 2, related work about color cues in object detection and recognition is reviewed. In Section 3, RMP is introduced and discussed. In Section 4 and Section 5, we carry out the experiments on visual object detection and object recognition. Section 6 concludes this paper.

2. Related work

In this paper, we do not separately review the work about color cues for both object detection and recognition, as features are utilized for both tasks with minor [29] or even without change [32,43].

Incorporating color cues into descriptors has received much more attention. In [36], the invariance and the distinctiveness of color descriptors are systematically studied for object recognition. One of the most successful color cues has been colored SIFT [31], where the color histogram and SIFT are concatenated into a vector. It is important to note that photometric invariant color space [31] is the primary step to obtain invariance, whereas coding schemes still confine to the simple statistics, *i.e.*, histogram or moment.

In practice, a successful system heavily relies on the combination of color, shape or texture. The classifier-level approach, such as MKL [32], boosting [8], pays much more attention on the choice of color features than that of designing descriptors, while the feature-level approach usually selects discriminative color channels to construct features [9,33,36]. Compared with the classifier-level one, the feature-level method results in more discriminative features, since it preserves the spatial binding between color and other cues.

For the feature-level approach, there are two important threads: one learns how to quantize the color space, a work based on topic modeling [36]; the other uses hand-crafted filters [33,35]. In the former case, the design of color descriptor involves modeling “color words” as Bag-of-Visual-Words (BoVW). For instance, [15] use probabilistic Latent Semantic Analysis (pLSA). However, they need to carefully tune the number of words [17].

On the contrary, hand-crafted filters totally depend on the predefined coding schemes. *E.g.*, Walk et al. use self-similarity to build Color Self-Similarity (CSS) [33]. However, CSS is the computation- and the memory-intensive descriptor. Compared with CSS, the proposed RMP produces more robust and compact features.

Recently Fisher Vector (FV) [26,27], and Convolutional Neural Network (CNN) [18] have demonstrated excellent performances for object detection and classification. FV parameterizes a probabilistic model from a hand-crafted feature space [27]. CNN requires a skillful training process and a large number of labeled examples. The purpose of this paper is to propose a hand-crafted feature which is expected to be generalize well without the complex learning process.

3. Rotative maximal pattern

3.1. The components of RMP

Fig. 1 illustrates that RMP consists of three steps: converting an image into patch-wise histograms, obtaining invariant responses from RCTs, and building RMP histogram.

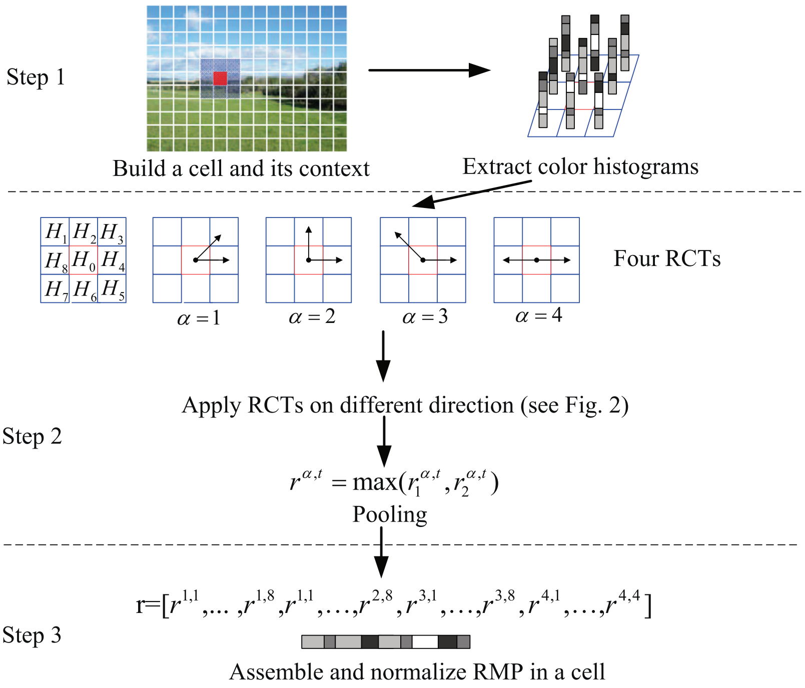
Step 1
Build a cell and its context
Extract color histograms
H_1 H_2 H_3
H_8 H_0 H_4
H_7 H_6 H_5
$\alpha = 1$
$\alpha = 2$
$\alpha = 3$
$\alpha = 4$
Four RCTs
Apply RCTs on different direction (see Fig. 2)
Step 2
$r^{\alpha,t} = \max(r_1^{\alpha,t}, r_2^{\alpha,t})$
Pooling
Step 3
$r=[r^{1,1},\ldots,r^{1,8},r^{1,1},\ldots,r^{2,8},r^{3,1},\ldots,r^{3,8},r^{4,1},\ldots,r^{4,4}]$
Assemble and normalize RMP in a cell

Fig. 1. The architecture of the feature extraction system proposed in this paper.

3.1.1. Step 1: convert images into color histograms

Our approach may be traced back to the early work on the patch-wise features [10]: building descriptors with features from image patches. Concretely, we first divide an image into non-overlapping cells, *e.g.*, 8 × 8 pixels regions, where a color histogram is extracted. The context of a cell is defined as the 3 × 3 cells (see Step 1 in Fig. 1). The invariance of a histogram is close to the choice of color spaces. The technical details of converting color spaces are beyond the scope of this work. We suggest to refer the literature [31] for the details. We empirically studied the influences of color spaces in Section 4.

Discussion. Histograms provide a certain robustness against deformation, translation and noise in a cell, while subdivided cells offset the potential loss of the spatial information. In such way, a compromise is achieved between the requirement of geometric invariance on the one hand, and that of the discriminative power on the other side. Thus, it is plausible that descriptors based on this compromise are simultaneously discriminative and robust, *e.g.*, the subregions in Spatial Pyramid Matching (SPM) [19], or SIFT [22].

3.1.2. Step 2: rotatable couple templates

To encode the different degrees of the rotations, we propose α-Couple Templates (α-CTs) where $\alpha \in \{1, 2, 3, 4\}$ index the possible rotations, *i.e.*, 45°, 90°, 135° and 180°, respectively (see Step 2 in Fig. 1). Given the 45° template as a example, Z_0, Z_3 and Z_4 form a local coordinate system with Z_0 as the original; thus, the angle between the template (Z_0, Z_3) and the one (Z_0, Z_4) is 45°.

The α-CTs are further rotated to capture more patterns in the 3 × 3 context, as illustrated in Fig. 2. Therefore, we extend the α-CTs into the $\alpha - t$ RCTs, where $t \in \{1, \ldots, 8\}$ index the possible rotations of the α-CTs (see Step 2 in Fig. 1). In nature, $\alpha - t$ RCTs serve as active templates to grasp the rotation-invariant patterns.

Let the dissimilarity function between the Z_0 patch and the Z_t one as $f_i^{\alpha}(H_0, H_t)$, the coupled responses in $\alpha - t$ RCTs, $\left(r_1^{\alpha,t}, r_2^{\alpha,t}\right)$, can be computed as follows:

$$r_1^{\alpha,t} = f_1^{\alpha}(H_0, H_t), \ \text{for } t = 1, \ldots, 8 \tag{1}$$

$$r_2^{\alpha,t} = f_2^{\alpha}\left(H_0, H_j\right), \ \text{for } j = (t + \alpha)\%8, \tag{2}$$

where the dissimilarity function $f(H_i, H_j)$ is defined as NHI,

$$f\left(H_i, H_j\right) = 1 - \frac{\sum_{k=1}^{K} \min(h_i^k, h_j^k)}{\max\left(\sum_{k=1}^{K} h_i^k, \sum_{k=1}^{K} h_j^k\right)}, \tag{3}$$

where h_i^k and k_j^k are the kth bin of the histograms H_i and H_j respectively, and K is the dimension of the histograms.

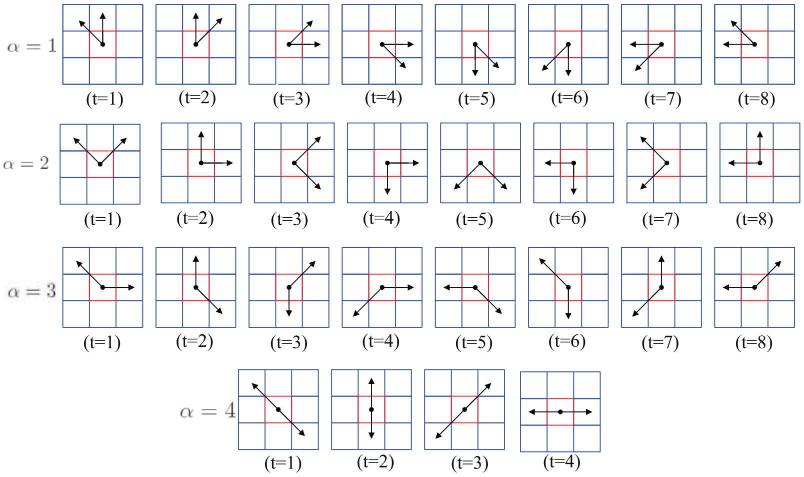

Fig. 2. The four $\alpha - t$ RCTs. (a), (b), (c) and (d) enumerate all possible rotations for 45°, 90°, 135° and 180°-CTs, respectively.

Intuitively, if $f(H_i, H_j)$ is larger than a threshold, it simulates the case that we attempt to capture the change of color distributions at the patch level. Moreover, NHI in (3) tolerates more noise as discovered in Theorem 1.

To achieve the invariance, max pooling is adopted on the coupled responses $(r_1^{\alpha,t}, r_2^{\alpha,t})$ as follows:

$$r^{\alpha,t} = \max(r_1^{\alpha,t}, r_2^{\alpha,t}). \tag{4}$$

The $\alpha - t$ RCTs can be considered as a special Receptive Field (RF) with a pooling operation [13]. This is a significant difference from the regular RFs [19,39] or overcompleted ones [14].

Discussion. The RCTs locally grasp the rotatable patterns at the different granularity. Therefore, the combination of the RCTs with the 45° rotation can not simply replace the 90°, the 135°, and the 180° ones. Average is the other straightforward yet widely used pooling strategy. Compared with max pooling, average pooling usually achieves more robust features than max pooling, but tends to loss the discriminative power [6].

3.1.3. Step 3: build a RMP histogram

Representing an image as the histogram of words has been used in the biologically plausible vision systems [2,22]. Motivated by this, as shown the Step 3 in Fig. 1, we vote the pooled responses $r^{\alpha,\,t}$ into RMP histogram for each cell. Denote RMP dictionary as $W^{\alpha,t} = \{w_1^{\alpha,t}, \ldots, w_m^{\alpha,t}, \ldots, w_{28}^{\alpha,t}\}$, where $w_m^{\alpha,t}$ corresponds the mth RCTs. Therefore, the mth bin of RMP is:

$$H(w_m^{\alpha,t}) = \delta(r_m^{\alpha,t}, w_m^{\alpha,t}) \cdot r_m^{\alpha,t}, \tag{5}$$

where δ is the Dirac delta function:

$$\delta(x, y) = \begin{cases} 0 & \text{for} \quad x \neq y \\ 1 & \text{for} \quad x = y. \end{cases}$$

Therefore, a three-dimensional RMP, $C_h \times C_w \times 28$, is generated for an image, if the image is divided into $C_h \times C_w$ cells. To obtain a more discriminative descriptor, the three-dimensional RMPs are further encoded into a feature vector: concatenating histograms from the 2 × 2 cells into a vector, and further normalizing it with ℓ_2 norm. A toy example is illustrated in Fig. 3.

3.2. Characteristics of RMP

RMP is not only robust to noise, but also is complementary to the pixel-wise features. RMP achieves invariance at the two levels, *i.e.*, voting operation of color histograms and RCTs with max pooling. Therefore, RMP is expected to be the complementary descriptor for other pixel-wise features, *e.g.*, Histogram of Gradient (HOG), and Local Binary Pattern (LBP). In addition, The edge-like patterns from RCTs correspond to the conclusion that much of the discriminative information is from the spatial frequencies [31]. Naturally, RMP is a powerful descriptor for both object detection and object recognition.

Next, we theoretically justify the robustness of RCTs. If one template of RCTs perfectly rotates into another one, the invariant response (1) is constantly selected by max pooling (4). On the contrary, if a template partially overlaps with the coupled one, one key question is why NHI (3) makes RCTs robust. Theorem 1 discovers the robustness of HI. The proof of Theorem 1 is given in Appendix.

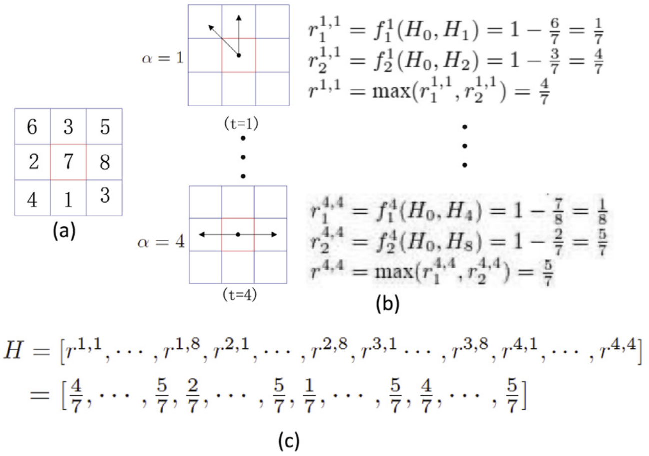
α = 1
$r_1^{1,1} = f_1^1(H_0, H_1) = 1 - \frac{6}{7} = \frac{1}{7}$
$r_2^{1,1} = f_2^1(H_0, H_2) = 1 - \frac{3}{7} = \frac{4}{7}$
$r^{1,1} = \max(r_1^{1,1}, r_2^{1,1}) = \frac{4}{7}$
(t=1)
6 3 5
2 7 8
4 1 3
(a)
α = 4
$r_1^{4,4} = f_1^4(H_0, H_4) = 1 - \frac{7}{8} = \frac{1}{8}$
$r_2^{4,4} = f_2^4(H_0, H_8) = 1 - \frac{2}{7} = \frac{5}{7}$
$r^{4,4} = \max(r_1^{4,4}, r_2^{4,4}) = \frac{5}{7}$
(t=4)
(b)
$H = [r^{1,1}, \cdots, r^{1,8}, r^{2,1}, \cdots, r^{2,8}, r^{3,1} \cdots, r^{3,8}, r^{4,1}, \cdots, r^{4,4}]$
$= [\frac{4}{7}, \cdots, \frac{5}{7}, \frac{2}{7}, \cdots, \frac{5}{7}, \frac{1}{7}, \cdots, \frac{5}{7}, \frac{4}{7}, \cdots, \frac{5}{7}]$
(c)

Fig. 3. A toy example of RMP. (a) An image patch represented by 1-dimensional histogram. (b) Compute responses from $\alpha - t$ RCTs; (c) Assembling the pooled responses into 28-dimensional RMP.

Table 1
A comparison among descriptors with our method for object detection and recognition.

Descriptors	Filtering	Labeling	Statistics
Color-SIFT	Intensity difference on color channels	Orientation quantization	Histogram over the weighted magnitudes in each channel
Color-LBP	difference on selected color space	Multi-scale directions	Histogram over both scale and directions
CSS	Histogram intersection	–	Histogram over the difference extracted among two cells
CN	Converting an image patch into histogram	quantizing histograms into words by pLSA	Histogram over the words
RMP	Histogram intersection	max pooling + RCTs	Histogram over RCTs

Theorem 1 (The robustness of HI)**.** *Given two histograms,* $H_i = [h_i^1, \ldots, h_i^K]$ *and* $H_j = [h_j^1, \ldots, h_j^K]$*, and let X be the random variable as* $X = \sum_{k=1}^{K} \min(h_i^k, h_j^k)$*. we further assume that* $\triangle = \max_k \triangle_k$ $(\triangle_k = \min(h_i^k, h_j^k))$*. Let E(X) be the expectation of the random variable X, and* ϵ *be a positive scalar* $\epsilon > 0$*, then we have:*

$$P(|X - E(X)| \geq \epsilon) \leq K\left(\frac{\triangle}{2\epsilon}\right)^2. \tag{6}$$

X can be considered as the degree of dissimilarity between two histograms. $|X - E(X)|$ naturally measures the robustness of HI for two histograms. To make HI more robust, we need to let the upper bound $\triangle$ as small as possible when K is fixed. To increase the robustness of RCTs, the normalization operation in (3) is used to reduce the bounds $\triangle$.

3.3. Comparisons with existing descriptors

Using the Filtering, Labeling and Statistic (FLS) framework, we easily compare RMP with the existing features in Table 1. In FLS framework, the filtering step depicts the inter-pixel relationship in a local region; the labeling step describes the redundancies including quantization and mapping; the statistic step captures the attributes in a local region. Note that our intent in this paper is of course *not* to compare the proposed descriptor to all features. We here instead wish to demonstrate that RMP can indeed meet or exceed these hand-crafted features on diverse tasks in both Sections 4 and 5.

For color-SIFT [31], it computes the magnitude and the orientation of gradients at each pixel from each color channel at the first stage. The orientations are then quantized to 8 dominant directions in the second stage. The gradient magnitudes are accumulated into an oriented histogram.

For color-LBP [44], it first converts an image into the predefined color spaces, and applies binary coding from LBP.

For CSS [33], it first encodes image patches by a color histogram, and computes the difference among patches. The difference between two patches is used to encode a feature vector.

For CN [36], it computes a color histogram from an image patch, and quantizes the histograms into 11 words by pLSA in the second stage. The words are further used to build the descriptor.

Although RMP also computes the dissimilarity between the center cell and its neighbors like LBP [24], the most stable responses are selected to build a feature vector. Different from color-SIFT [31] and color-LBP [44], RMP uses histograms from cells to build the feature vector. Different from CSS [33], RMP utilizes RCTs to obtain the invariance ability.

4. Application to object detection

4.1. RMP for object detection and the data sets

Recently, many approaches, especially for pedestrian [8,33], have been proposed for object detection. Among these methods, the sliding window approach has attracted much attention, such as [7,11]. In this paper, RMP is applied for object detection by the sliding window approach.

The experiments are carried out on two challenge data sets: INRIA pedestrian [7] and Pascal VOC2007 [10]. INRIA pedestrian consists of 1218 negative training photos and 2478 positive samples with left-right reflections. By adopting the protocol of INRIA pedestrian, in the first round training, 12,180 negative samples from 1218 negative training photos and 2478 positive ones are used to train a classifier. Then an augmented set (12,180+hard samples + 2478 positive examples) produces the final detector. Some examples of INRIA Person are illustrated in Fig. 7.

Pascal VOC2007 consists of 9963 images from 20 classes with 5011 training images and 4952 test ones. The Pascal VOC2007 is considered as the extremely challenging ones because all the images are daily photos where the size, viewing angle, illumination etc appearances of objects are vary significantly with frequent occlusions.

4.2. Training detectors and baselines

At the training stage, one of the core problem in the sliding window approach is how to combine multiple features. The color and shape descriptors can be fused at the two stages: before and after the alignment of examples. Therefore, we empirically study different fusion strategies as follows:

- *Before fusion:* Concatenate HOG and RMP into a new feature for a detector;
- *After fusion:* First use HOG to align examples, and then concatenates HOG and RMP into a new feature to retrain a detector.

Two standard detection approaches are used in our experiments. In the holistic approach, HOG is extracted densely from a 8 × 8 pixels cell, and a 28 dimensional RMP histogram is encoded from the cell. In the part-based approach [11], each object is modeled as a deformable collection of parts. Baselines [7], [33], and [15] are summarized in the subsequent comparisons:

1. HOG [7]: We compare it to find whether RMP is complementary to the shape-based feature or not.
2. (HOG+CSS) [33]: [33] reports the state-of-the-art performances on INRIA Pedestrian by combining HOG and the color-based feature.
3. (HOG+CN) [15]: We further compare RMP with a recent work [15]. Different from our approach, CN clusters RGB-based histograms into 11 color words by pLSA.

In the evaluation stage, Non-Maximal Suppression (NMS) is used. A candidate is recognized as a successful detection, only if the condition for the detected box and the groundtruth satisfies that $\frac{\text{intersection}}{\text{union}} > 0.5$. The miss rate against False Positive Per Image (FPPI) is used to evaluate performances.

4.3. The experimental results

4.3.1. Parameter choices for RMP

There are three key parameters in RMP:

- the choice of a color space,
- the pooling strategy in RMP,
- the dissimilarity function.

To leverage the photometric invariance from a color histogram, we compute a color histogram from each cell and use the trilinear interpolation to minimize the aliasing problem. More concretely, each color space is uniformly quantized into 9 bins, and thus the color histogram for a block is encoded as a 9 × 3 × 4 histogram with ℓ_2 normalization. In Fig. 4(a), 12 different color spaces are compared under the same experimental settings.

As shown in Fig. 4(a), compared with the normalized HS, the color histogram from Lab space obtains a significant gain of 8% at 1 FPPI. According to the conclusion that the human eyes have three different types of color sensitive cones [23], the intensity ignored by normalized color spaces is critical to describe objects. Naturally, the performances of normalized RG, normalized HS and normalized UV [33] are worse than RGB, HSV, and YUV. In summary, the top-4 color spaces are ranked in a descend order, *i.e.*, Lab, LUV, HSV, and HSI.

Fig. 4(b) further compares the effectiveness of RMP with Lab, LUV, HSV, and HSI. RMP achieves similar performances among these color spaces, *i.e.*, around 0.30 miss rate at 1 FPPI, in LUV, Lab, and HSV. Compared with the results from color histograms in Fig. 4(a), RMP approximately provides an improvement of 20% accuracies over color histograms. Note that both HOG and LBP are used to replace color histograms, but are inferior to color histograms in our unreported experiments.

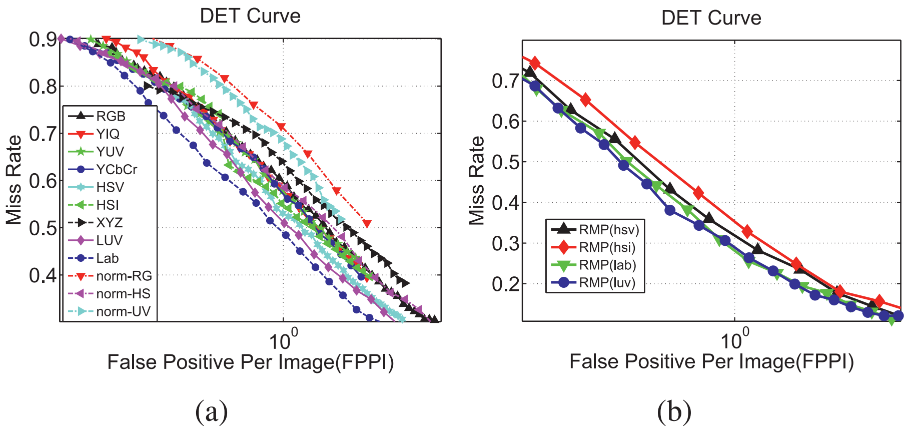

Fig. 4. The detection results as a function of color spaces with color histograms (a) and with RMP (b) on INRIA pedestrian.

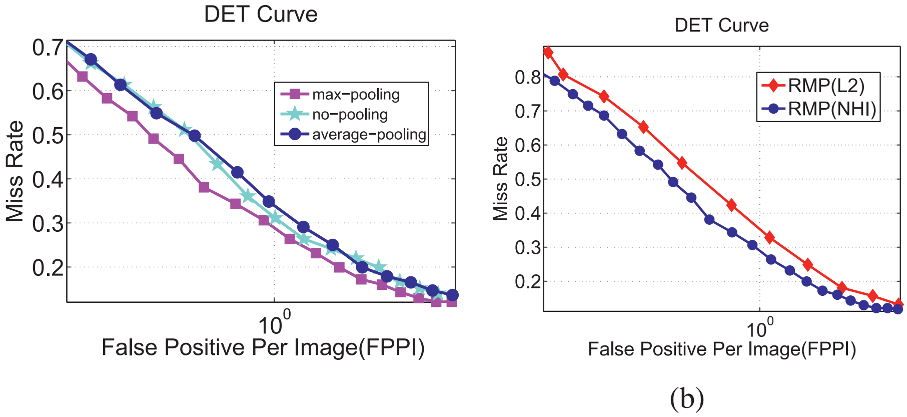

Fig. 5. The detection results as a function of the pooling methods (a), and the similarity functions (b) on INRIA pedestrian data set.

The effectiveness of pooling strategy is shown in Fig. 5(a). Max pooling outperforms both no-pooling and average pooling. As expected, the performances of average pooling are slightly worse than that of no-pooling. This indicates that average pooling indeed damages the discriminative power of features.

Finally, the choice of similarity functions is discussed. Fig. 5(b) compares NHI and Euclidean distance. As expected in Theorem 1, NHI outperforms Euclidean distance by about 4% accuracy at 1 FPPI. The result indicates that NHI is robust to distortion caused by rotations. This is critical to both object detection and recognition, since rotation is one of the typical variations.

In conclusion, our experimental results suggest that RMP consistently provides improved performances over color histograms. Besides, RCTs, max pooling, and NHI comprise the discriminative yet invariant power for RMP.

4.4. Comparative evaluation

In this section, we compare the proposed approach to the state-of-the-art hand-crafted features on INRIA pedestrian and Pascal VOC2007. To make the comparisons as meaningful as possible, we follow the protocols defined by each data set.

4.4.1. Experiments on INRIA data set

For INRIA pedestrian, one round of boostrapping procedure is adopted [7]. As INRIA pedestrian already supplies the aligned examples, *before fusion* is adopted to combine HOG and RMP.

We begin with an analysis of the computational costs between CSS and RMP. Theoretically, if an image is divided into $W \times H$ cells, the computational costs of CSS and RMP are $O((WH-1)!)$ and $O(8 \times (W-1)(H-1))$, respectively. For example, for a 128 × 64 pixels image, the dimension of RMP is 3,072, while the dimension of CSS is 8,128. Empirically, on a 3.3GHz CPU with 8 GB memory under Matlab, CSS needed about 1351 seconds while RMP only required around 538 seconds for 1000 samples from INRIA data set. That is, RMP has a lower computational cost than CSS.

Fig. 6(a) shows that the performances of RMP are better than that of CSS with normalized SV [33]. The CSS yields a 0.44 miss rate at 1 FPPI, while the best result obtained by RMP(Lab) provides an improvement of 16% over CSS at 1 FPPI.

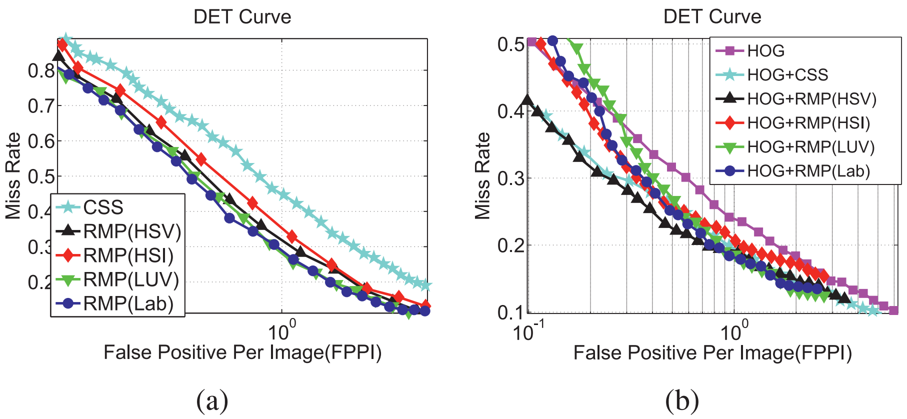

Fig. 6. Results from RMP, CSS, and their combination with HOG on INRIA pedestrian. (a) Results from RMP with different color spaces and CSS with normalized HV. (b) Results between RMP and CSS with the combination of HOG.

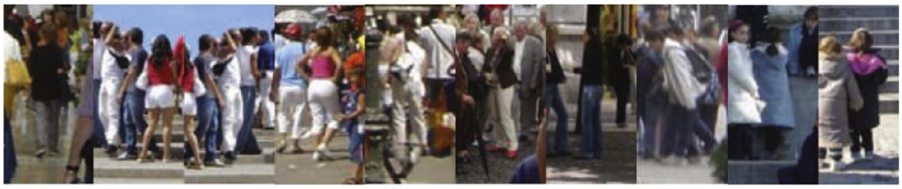

Fig. 7. The false negatives generated by HOG but successfully detected by HOG-RMP(HSV).

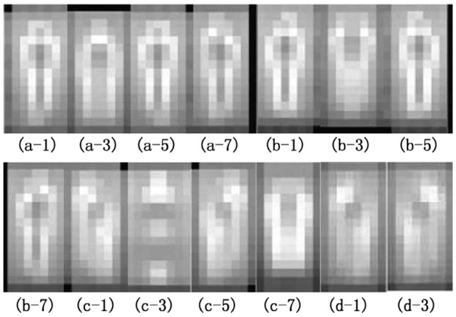

Fig. 8. The feature maps on INRIA data set. The index of a map corresponds to the one in Fig. 2.

Fig. 6(b) further illustrates the combination of HOG and RMP to understand whether the descriptors complement each other or not. HOG-RMP(HSV) outperforms HOG by 7% accuracy at 1 FPPI and by 10% one at 0.1 FPPI. As a comparison, HOG-RMP(HSV) outperforms HOG-CSS(norm-SV) by 2% accuracy at 1 FPPI. Fig. 7 shows some false negatives generated by HOG, but successfully detected by HOG-RMP(HSV). It can be observed that HOG-RMP(HSV) handles complex backgrounds, pose variations and photometric variance.

Fig. 8 vividly illustrates some feature maps of RMP for INRIA data set. There are some redundancies between HOG and RMP. As a result, RMP(HSV) achieves 0.18 missing rate over CSS(norm-SV) at 1 FPPI in Fig. 6(b), while HOG-RMP(HSV) only outperforms HOG-CSS(norm-SV) by 2% at 1 FPPI in Fig. 6. RMP, a patch-wise feature, captures the coarse boundaries of objects. Therefore, RMP can be still considered as a complementary feature for the pixel-wise descriptor.

In order to verify the invariance of RMP, Fig. 9 compares the robustness of both CSS and RMP when different sliding steps are applied. When the steps are changed, the performances of RMP decline much slowly than that of CSS. Specially, the performances of RMP nearly do not change when the sliding steps are from 8 to 12 (see details in Table 4).

In summary, compared with CSS, RMP, with a lower computational cost, achieves a better discriminative ability; besides, RMP is robust to translation and noise.

4.4.2. Experiments on VOC 2007

We first analyze the abilities of different feature fusion strategies. Table 2 gives a per-category performances between *before fusion* and *after fusion*. Firstly, except “cow”, both *before fusion* and *after fusion* improve performances over HOG alone.

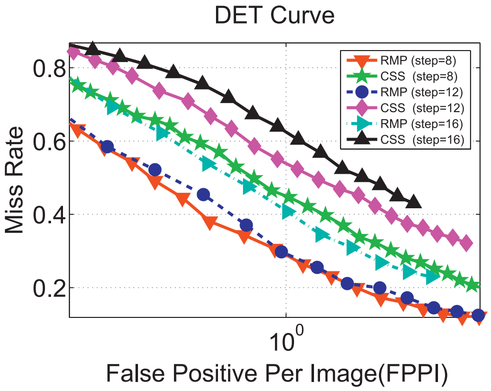

Fig. 9. The comparison between CSS and RMP when different sliding steps are used on INRIA pedestrian data set.

After fusion slightly outperforms *before fusion*, except the objects, "bus", "horse", and "person". This indicates that the alignment tends to improve performances.

Recently, the comprehensive evaluation of color descriptors [15] shows that CN achieves a superior performance on Pascal VOC2007. In this experiment, CN and other color descriptors are considered as baselines, although RMP is not a color descriptor but utilizes the photometric invariance from colors. Since a recognition system involves many aspects, such as tuning parameters, we directly copy the reported results from the original papers [15,32,42].

Table 3 shows the results on all 20 categories of Pascal VOC2007. Note that the results of RGB-HOG, OPP-HOG, C-HOG and CN-HOG are directly cited from the literature [15]. None of the three color descriptors, *i.e.*, RGB-HOG, OPP-HOG, and C-HOG, improves the performances over HOG. Compared with HOG, CN-HOG significantly improves the accuracies, especially for "sheep", "plant", "horse", "cat", "boat" and "aero". Interestingly, CN-HOG simultaneously decreases the performances of the other 5 objects, *i.e.*, "bottle", "bus", "chair", "person", "sofa". As a contrast, RMP always increases the performances of HOG. The reason is that RMP is a shape-like feature, while CN is a pure color feature. It also verifies the generalization ability of RMP across different objects.

Table 3 also shows the comparisons between RMP and the state-of-the-art hand-crafted features. Oxford-MKL [32] uses the BoVW framework with multiple weighted features. LEO [45] uses a latent hierarchical structural model to train a object detector. LBP-HOG [42] combines HOG and LBP by boosting. It should be noted that without the feature selection procedure, LBP-HOG provides inferior results. HOG-RMP(HSV)-A performs slightly worse than LBP-HOG. In contrast to LBP-HOG, no feature selection procedure is used in HOG-RMP(HSV)-A, though a selection strategy is easily incorporated into RMP which is expected to further improve performances. Note that the results of OXford-MKL [32], LBP-HOG [42], LEO [45] are directly copied from their papers. (Table 4).

In summary, the best combination of RMP and HOG is achieved when *after fusion* is adopted for object detection. Moreover, RMP has a good generalization ability.

5. Application to object classification

5.1. Background

Object classification with the hand-crated features usually uses BoVW to represent an image. One of the successful BoVW is SPM, which partitions an image into several subregions,*e.g.*, 1 × 1, 2 × 2, and 3 × 3. Firstly, the extracted local descriptors are converted into a dictionary, and then the extracted features are voted into a feature vector.

5.2. Data sets and baselines

- Caltech101 involves 102 object categories, where each category contains 31 to 800 images (see Fig. 10(a)). We resize these images to be less than 300 × 300 pixels with the aspect ratio kept unchanged. We follow the common setup where 15 and 30 training images per category are tested. The number of testing samples for each object category is fixed at 15.
- Scene15 contains 15 scene categories, where each category has 200–400 positive samples with an average size of 300 × 250. The major sources of pictures in this set include the COREL collection, personal photographs and Google Image Search. The total image number is 4485 (see Fig. 10(b)). Following the same experiment setup by [19], 100 positive training samples are randomly selected for each category and the rest is for testing.
- UIUCsport contains eight categories with 1792 images. The eight categories are: badmintion, bocce, croquet, polt, rock climbing, rowing, sailing, and snow boarding. The number of images ranges from 137 to 250 (see Fig. 10(c)). The UIUC-sport is designed to test the where-who-what problem [21].

Table 2
Comparisons between before fusion and later fusion on VOC 2007.

	aero	bicyc	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [11]	28.9	59.5	10.0	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	32.3
HOG-RMP-B	31.9	60.2	10.9	16.8	25.9	51.8	58.3	20.8	23.6	25.1	24.4	11.8	59.0	48.7	43.5	13.6	19.8	35.1	46.7	43.2	33.5
HOG-RMP-A	32.6	61.2	11.3	17.6	26.4	51.2	58.7	21.4	23.8	25.1	25.6	12.1	58.4	49.2	43.1	13.8	20.4	35.6	47.5	45.1	34.0

HOG-RMP-B indicates the before fusion scheme while HOG-RMP-A represents the after fusion one.

Table 3
Average precision results for object detection on Pascal VOC 2007.

	aero	bicyc	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [11]	28.9	59.5	10.0	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	32.3
CN-HOG	34.5	61.1	11.5	19.0	22.2	46.5	58.9	24.7	21.7	25.1	27.1	13.0	59.7	51.6	44.0	19.2	24.4	33.1	48.4	49.7	34.8
OPP-HOG	29.2	54.2	10.7	14.5	17.9	45.8	53.5	21.7	19.3	22.8	21.7	12.3	57.4	46.0	41.2	15.6	19.2	25.0	42.2	41.2	30.6
C-HOG	29.1	54.7	9.8	14.3	17.9	44.8	55.2	16.0	19.5	25.1	19.6	11.8	58.5	46.6	27.1	15.2	19.0	26.9	44.0	46.6	30.1
RGB-HOG	33.9	56.5	6.8	13.7	22.9	46.2	56.6	14.9	20.4	22.8	19.3	11.7	57.1	46.7	40.6	13.3	19.2	31.6	47.5	43.4	31.3
Oxford-MKL	37.6	47.8	15.3	15.3	21.9	50.7	50.6	30.0	17.3	33.0	22.5	21.5	51.2	45.5	23.3	12.4	23.9	28.5	45.3	48.5	32.1
LBP-HOG	36.7	59.8	11.8	17.5	26.3	49.8	58.2	24.0	22.9	27.0	24.3	15.2	58.2	49.2	44.6	13.5	21.4	34.9	47.5	42.3	34.3
LEO	29.4	55.8	9.4	14.3	28.6	44.0	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	29.6
HOG-RMP(HSV)-A	32.6	61.2	11.3	17.6	26.4	51.2	58.7	21.4	23.8	25.1	25.6	12.1	58.4	49.2	43.1	13.8	20.4	35.6	47.5	45.1	34.0

Table 4
Comparison between CSS and RMP on INRIA pedestrian at 1 FPPI (Miss rate %).

Feature	Dim.	pooling	Step = 8	Step = 12	Step = 16
RMP	3072	Max pooling	29.5	29.6	41.5
CSS	8128	–	42.9	51.8	60.4

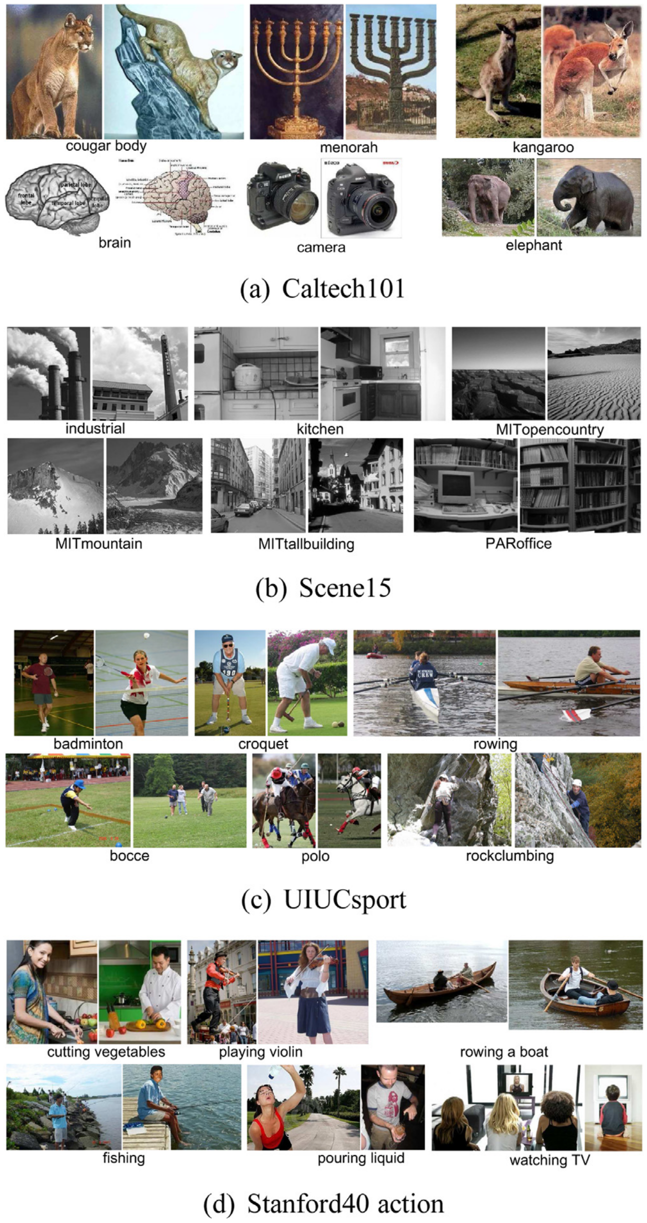

Fig. 10. Example images from different data sets. The coupled images belongs to the same category.

- Stanford40 action is one of the most challenging action recognition data sets. Stanford40 consists of 9532 images from 40 action categories, such as jumping, repairing a car, cooking, applauding, cutting vegetables, throwing a fishbee. The large number of action categories makes this data set particularly challenging (see Fig. 10(d)).

These data sets used here are very diverse, in order to evaluate the generalization ability of RMP. During experiments, we follow the experimental setups proposed by the different designers. For all data sets, a 3-level pyramid is used with the grid size 4 × 4, 2 × 2 and 1 × 1, yielding a total of 21 pyramid cells. To fusion different types of features, the *before fusion* is used. For classification, we use a nonlinear SVM with a χ^2 kernel.

Table 5
Classification error rates (%) on Caltech-101(Rec. in %).

Algorithms	15 training	30 training
SVM+LLC [34]	65.4	73.4
MKL(GB-PHOG$_{g/c}$-SS) [32]	71.1± 0.6	78.2± 0.4
SVM(RMP)	57.4± 0.4	63.5± 0.8
SVM(SS)	57.2± 0.2	62.2± 0.6
SVM(PHOG$_g$)	55.3± 1.1	62.2± 1.4
SVM(SIFT)	60.3± 0.8	63.2± 1.3
SVM(PHOG$_g$-SS-SIFT)	67.3± 0.6	73.5± 0.8
SVM(RMP-PHOG$_g$-SS-SIFT)	72.5± 0.8	79.4± 0.5

Table 6
Performance comparisons on the Scene15(Rec. in %).

Algorithms	Performance
KSPM [19]	81.4± 0.5
ScSPM [39]	80.3± 0.9
KD [4]	86.7± 0.4
GIST [25]	69.5
BED [38]	84.1± 0.5
SVM(RMP)	66.7± 0.4
SVM(SS)	58.5± 0.5
SVM(PHOG$_g$)	77.8± 0.5
SVM(SIFT)	80.5± 0.3
SVM(PHOG$_g$-SS-SIFT)	83.7± 0.5
SVM(RMP-PHOG$_g$-SS-SIFT)	87.2± 0.3

5.3. Experimental results on diverse data sets

Here we present results of RMP and its combination with other descriptors. Dense sampling is used to extract descriptors from image regions. Several descriptors are involved in our experiments. One local appearance feature (SIFT [22]), two shape features (SS and Pyramid Histogram of Oriented Gradient (PHOG) from both grey and HSV space, namely PHOG$_g$ and PHOG$_c$ respectively) and RMP are used. In particular, SIFT is computed in grey space over a square patch of radius with the spacing of r. We take $r = 4$, 8 and 12 pixels to allow scalability. SS descriptor is used to capture a correlation map of a 5 × 5 patch with its neighbors at every 5th pixel, and the correlation map is quantized into 10 orientations and 3 radial bins to form a 30 dimension descriptor. PHOG, in this paper, employs 30 orientation bins in 360°. To give a fair comparison with SS, RMP also uses the 5 × 5 pixels to extract the patch-wise descriptor. We employ k-means to quantize these descriptors to obtain dictionaries of size k (say, 1200), respectively.

5.3.1. Results on Caltech101

Table 5 compares the proposed method with several other methods on Caltech101. For Caltech101, the following baselines are compared: 1) MKL [32] combines the geometric blur (GB) [3,41], PHOG$_g$, PHOG$_c$ and SS; 2) SVM+LLC [34] uses the local reconstruction coefficients to code every SIFT. The results of the baselines [32,34] are directly copied from the reported papers.

First, our combined features achieve the state-of-the-art performance. Besides, RMP slightly outperforms SS when the number of training sample is 30. As illustrated in Fig. 10(a), most of objects from Caltech101 not only live at the center of images, but also have a similar scale. Therefore, the invariance ability of RMP does not be fully examined on this data set.

5.3.2. Results on Scene15

Table 6 reports comparisons between RMP and several methods on Scene15. Kernel SPM (KSPM) [19], Sparse Coding for SPM (ScSPM) [39], kernel descriptor (KD) [4], and Beyond Euclidean Distance (BED) [38] are selected as baselines. KSPM uses SIFT to generate SPM for the kernel SVM. ScSPM uses sparse coding to handle the quantization problem in BoVW. KD converts the pixel attributes into patch-wise features by match kernels. BED uses histogram intersection kernel to generate a codebook. GIST [25] is specially designed for scene recognition, with a set of perceptual dimensions, *i.e.*, naturalness, openness, roughness, expansion and ruggedness. KSPM, ScSPM and BED all focus on the generation of an effective codebook. Noted that the results of all these baselines [4,19,25,38,39] are directly cited from their papers.

As illustrated in Table 6, the combined features outperform ScSPM by about 7% and KSPM by 5.8%, respectively. Intuitively, Scene15 requires a feature not only to describe the texture-like objects (*e.g.*, “MITopencountry”, “MITmoutain”), but also to grasp the shape-like ones (*e.g.*, “industrial”, “kitchen”).

Fig. 11 further shows the confusion matrix among the 15 scene categories. Top-4 most confused classes occur among “livingroom”, “industrial”, “bedroom” and “insidecity”. As expected, RMP, the patch-wise feature, is difficult to distinguish the details between similar categories, *e.g.*, “livingroom” and “bedroom”. However, RMP still outperforms SS by about 8%

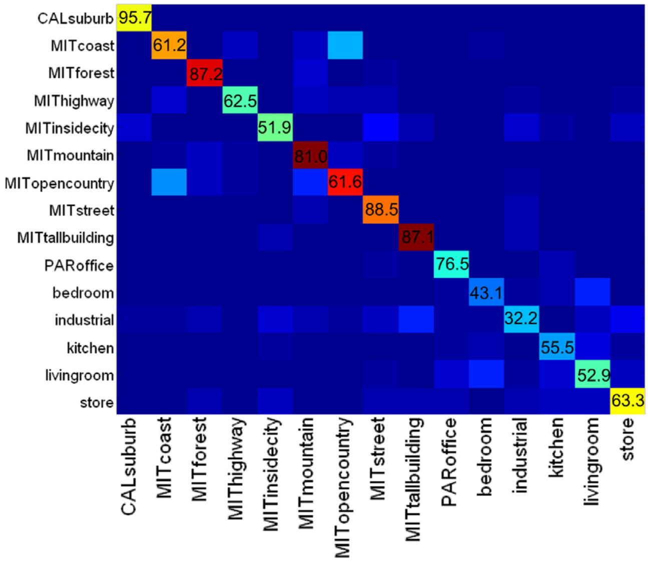

Fig. 11. Confusion matrix of RMP on Scene15. Average accuracy rates for each class are listed along the diagonal.

Table 7
Performance comparisons on the UIUC sport(Rec. in %).

Algorithms	Performance
ScSPM [39]	82.7 ± 1.5
GIST [25]	70.7 ± 0.6
LscSPM [12]	85.3 ± 0.5
OB [20]	77.8
SVM(RMP)	75.2 ± 0.6
SVM(SS)	73.3 ± 0.8
SVM(PHOG$_g$)	59.0 ± 1.2
SVM(SIFT)	77.5 ± 1.5
SVM(PHOG$_g$-SS-SIFT)	83.1 ± 0.8
SVM(RMP-PHOG$_g$-SS-SIFT)	88.5 ± 0.7

accuracy. This indicates that the combination of RCTs and max pooling indeed grasps more discriminative patterns than the simple dissimilarity in SS. Although KD outperforms RMP by 20% accuracy, KD is computationally intensive since it projects every pixel into a kernel space. Naturally GIST outperforms RMP by about 3%, because GIST is specially designed for scene recognition tasks.

5.3.3. *Results on UIUCsport*

Laplacian Sparse Coding for SPM (LscSPM) [12], ScSPM [39], GIST [25], and Object Bank (OB) [20] are considered as baselines. LscSPM [12] achieves a better performance than ScSPM by adding Laplacian constraint. OB [20] converts the low-level features into the semantic ones by object detectors. The results of these baselines [12,20,39] are directly cited from their papers.

Table 7 illustrates that the patch-wise features, both RMP and SS, perform better than the pixel-wise one, *i.e.*, PHOG$_g$. Compared with OB, RMP achieves a comparable result. Note that OB needs a set of pretrained detectors which indirectly supply some side information for visual tasks. Moreover, RMP outperforms SS by about 2% and GIST by about 5%, respectively. As discussed in [21], the categories in UIUCsport mostly are scene related. The context tends to supply discriminative cues, explaining why RMP consistently outperforms SS on both Scene15 and UIUCsport. (Table 8).

5.3.4. *Results on Standford40 Action*

We first summarize our baselines on this data set, Human Attributes (HA) [40], Expanded Parts Model (EPM) [30], and OB [20], CN [16] and RG-SIFT [16]. HA [40] requires a set of pretrained poselets [5] and classifiers. EPM [30] decomposes human pose into a set of parts, and uses the “immediate” and “full image” context. RG-SIFT [16] combines the R, G color channels into SIFT. For Standford40, the interactive actions require features not only to describe scenes, but also to distinguish different objects. The results of these baselines [16,30,40] are directly cited from their papers.

OB, HA and EPM all obtain inferior performances, although these methods try to build the middle- and even high-level descriptions for actions. RMP outperforms CN by 10% and SS by 4%. It means that dissimilarity pattern is more discriminative than the color distribution for action recognition. RMP consistently performs well on objects classification (on Caltech101),

Table 8
Performance comparisons on the stanford 40 action(Rec. in %).

Algorithms	Performance
SVM(CN) [16]	17.6
EPM+context [30]	42.2
HA [40]	45.7
SVM(RG-SIFT) [16]	39.6
OB [20]	32.5
SVM(RMP)	27.6
SVM(SS)	23.9
SVM(PHOG_g)	39.1
SVM(SIFT)	38.2
SVM(PHOG_g-SS-SIFT)	43.5
SVM(RMP-PHOG_g-SS-SIFT)	46.3

Table 9
Comparisons among the size of cells in RMP (Rec. in %).

Size	Caltech101_{15}	Scene15	UIUCsport	Stanford40
$\text{RMP}_{3\times3}$	59.2	71.3	75.3	29.5
$\text{RMP}_{5\times5}$	57.4	68.2	74.6	27.6
$\text{RMP}_{7\times7}$	52.7	61.6	72.8	23.4

Table 10
Performance comparisons on feature combination (Rec. in %).

Methods	Caltech101_{15}	Scene15	UIUCsport	Stanford40
$\text{RMP}_{5\times5}$-SS	64.3 (+6.0)	71.2 (+12.7)	79.8 (+6.5)	33.3 (+9.4)
$\text{RMP}_{5\times5}$-PHOG_g	68.7 (+13.4)	78.6 (+0.8)	82.5 (+23.5)	40.2 (+1.1)
$\text{RMP}_{5\times5}$-SIFT	69.5 (+9.2)	82.3 (+1.8)	82.7 (+5.2)	40.7 (+2.5)

scene recognition (on Scene15), and scene-object recognition (on UIUCsport). The performances on Stanford40 demonstrate that RMP has a good generalization ability across different visual tasks.

In summary, RMP outperforms the patch-wise features, *i.e.*, SS and CN, and further shows a good generalization ability across different tasks. As mentioned before, a certain degree of invariance from both RCTs and max pooling empowers RMP a higher discriminative power than other descriptors.

5.3.5. Experiments revisit

We empirically show two important factors of RMP: 1) the size of a cell; and 2) the complementary ability for different features.

Table 9 shows that smaller the size of a cell is, a better the performance RMP achieves. Especially for Scene15, $\text{RMP}_{5\times5}$ outperforms $\text{RMP}_{7\times7}$ by about 7% accuracy. It indicates that RMP grasps more discriminative patterns in a smaller region.

Moreover, Table 10 further shows the complementary ability of RMP with other descriptors. The positive numbers in the brackets mean the relative performance improvements. For example, “64.3(+6.0)” means the combined feature achieves 64.3% accuracy and an improvement 6.0% over SS alone. There are two observations from Table 10:

- The complementary ability of RMP depends on both descriptors and tasks. For instance, $\text{RMP}_{5\times5}$-PHOG_g obtains 23.5% improvement over PHOG_g on UIUCsport, but nearly has not gain on Scene15.
- $\text{RMP}_{5\times5}$-SS consistently outperforms SS by at least 6% on all data sets. It again demonstrates that RCTs with max pooling indeed grasps more discriminative patterns than SS.

6. Conclusion and future work

In this paper, RMP is proposed to fuse color and shape cues into a local coloring descriptor. It is invariant, discriminative, and complementary to pixel-wise descriptors. The invariance comes from three aspects: the photometric invariance from colors, max pooling of RCTs, and the dissimilarity from NHI. The discriminative power comes from the shape-like pattern encoded by RCTs. The complementary ability mainly comes form the patch-wise features build from color histograms. Experiments show that RMP achieves promising performances on INRIA pedestrian and Pascal VOC2007 for the visual detection task. On object classification, with comparison to the state-of-the-art hand-crafted features [15,20,32], experiments demonstrate that RMP achieves comparable or exceeding performances on Scene15, Caltech101, UIUCsport and Standford40 action. Although these visual tasks and the corresponding data sets are very diverse and different, RMP exhibits a good generalization ability.

The promising results of this paper motivates a further examining of RCTs-based descriptors: First, the use of a larger context in RMP, like 4 × 4, may provide more discriminative cues over the 3 × 3 used here. Furthermore, the proposed approach can be embedded into the learning based framework [28].

Acknowledgement

This work was supported in part by Natural Science Foundation of China under Grant 61672069, Grant 61332016, Grant 61620106009, Grant 61472387, Grant U1636214, and Grant 61650202, in part by Key Research Program of Frontier Sciences under Grant CAS: QYZDJ-SSW-SYS013, in part by China Postdoctoral Science Foundation, by Beijing Post-Doctoral Research Foundation, by Beijing Municipal Commission of Education under Grant KM201610005034, and by the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR).

Appendix

We prove Theorem 1 here.

Proof. According to Chebyshev's inequality, if $\forall \epsilon > 0$, we have:

$$P(|X - E(X)| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}, \tag{7}$$

where $Var(X)$ is the variance of the random variable X. Let $T_k = \min(h_k^i, h_k^j)$ be Bernoulli random variable, as the probability of the event $h_k^i \neq h_k^j$ is p, $(0 \leq p \leq 1)$. If the variable $T_k, k = 1, \ldots, K$ are independent to each others, the variance $Var(X)$ can be computed as:

$$Var(X) = \sum_{k=1}^{K} Var(T_k). \tag{8}$$

The variance of T_k, $Var(T_k)$, is further calculated as:

$$\begin{aligned} Var(x_k) &= E(T_k^2) - (E(T_k))^2 \\ &= 0(1-p) + \triangle_k^2 p + (0(1-p) + \triangle_k p)^2 \\ &= \triangle_k^2 p - \triangle_k^2 p^2 \\ &= \triangle_k^2 p(1-p) \end{aligned} \tag{9}$$

Therefore, putting (9) into the inequality (7), we have

$$\begin{aligned} &P(|X - E(X)| \geq \epsilon) \leq \tfrac{K\triangle_k^2 p(1-p)}{\epsilon^2} \\ &\leq \tfrac{K\triangle^2 p(1-p)}{\epsilon^2} \\ &\leq \tfrac{K\triangle^2}{4\epsilon^2} (\text{when} \quad p = \tfrac{1}{2}) \end{aligned} \tag{10}$$

□

References

[1] A.E. Abdel-Hakim, A.A. Farag, Csift: a sift descriptor with color invariant characteristics, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 3, 2006, pp. 1978–1983.

[2] M. Agrawal, K. Konolige, M. Blas, Censure: center surround extremas for realtime feature detection and matching, in: Proc. European Conf.on Computer Vision, 4, 2008, pp. 102–115.

[3] A. Berg, T. Berg, J. Malik, Shape matching and object recognition using low distortion correspondences, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 2126–2136.

[4] L. Bo, X. Ren, D. Fox, Kernel descriptors for visual recogntion, in: Advances in Neural Information Processing Systems, 1, 2010, pp. 244–252.

[5] L. Bourdev, J. Malik, Poselets: body part detectors trained using 3d human pose annotations, in: International Conference on Computer Vision, 3, 2009, pp. 1365–1372.

[6] Y. Boureau, J. Ponce, A theoretical analysis of feature pooling in visual recognition, in: International Conference on Machine Learning, 2, 2010, pp. 111–118.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 886–893.

[8] P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral channel features, in: British Machine Vision Conference, 2009, pp. 1–11.

[9] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pdestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761.

[10] M. Everingham, L. Gool, C. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[11] P. Felzenszwalb, D. McAllester, D.Ramanan, A discriminative trained, multiscale, deformable part model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2, 2008, pp. 2126–2134.

[12] S. Gao, I. Tsang, L. Chia, P. Zhao, Local feature are not lonely: Laplacian sparse coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 4, 2010, pp. 3555–3561.

[13] D. Hubel, T. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (2) (1962) 106–154.

[14] Y. Jia, C. Huang, T. Darrell, Beyond spatial pyramids: receptive field learning for pooled image features, in: IEEE Conference on Computer Vision and Pattern Recognition, 4, 2012, pp. 3370–3377.

[15] F. Khan, R.M. Anwer, J. Weijer, A. Bagdanov, M. Vanrell, A. Lopez, Color attributes for object detection, in: IEEE Conference on Computer Vision and Pattern recognition, 1, 2012, pp. 92–100.

[16] F. Khan, R.M. Anwer, J. Weijer, A.D. Bagdanov, A.M. Lpez, M. Felsberg, Coloring action recognition in still images, Int. J. Comput. Vis. 24 (7) (2014) 971–987.

[17] R. Khan, J.V. Weijer, F.S. Khan, D. Muselet, C. Ducottet, C. Varat, Discriminative color descriptors, in: IEEE Conference on Computer Vision on Computer Vision, 2, 2013, pp. 14–20.

[18] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Neural Information Processing Systerms, 2012, pp. 1106–1114.

[19] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Commpute Vision and Pattern Recognition, 2, 2006, pp. 2169–2178.

[20] L. Li, E.P. Xing, F.-F. Li, Object kank: a high-level image representation for scene classification and semantic feature sparsification, in: Advances in Neural Information Processing Systems, 1, 2010, pp. 630–637.

[21] L.J. Li, F.-F. Li, What, where and who? Classifying events by scene and object recogintion, in: International Conference on Computer Vision, 2, 2007, pp. 14–20.

[22] D. Lowe, Distinctive image feature from scale invariant key points, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[23] J. Ng, A. Bharach, Z. Li, A survey of architecture and function of the primary visual cortex, in: Eurasip Journal on Advances in Signal Processing, 1, 2007, pp. 92–100.

[24] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[25] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.

[26] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: theory, and practice, Int. J. Comput. Vis. 105 (3) (2013) 222–245.

[27] K. Sande, C. Snoek, A. Smeulders, Fisher and vlad with flair, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 23–28.

[28] L. Shao, L. Liu, X. Li, Feature learning for image classification via multiobjective genetic programming, IEEE Tans. Neural Netw. Learn. Syst. 25 (7) (2014) 1359–1371.

[29] L. Shao, X. Zhen, D. Tao, X. Li, Spatio-temporal laplacian pyramid coding for action recognition, IEEE Tans. Cybern. 26 (1) (2014) 817–827.

[30] G. Sharma, F. Jurie, C. Schmid, Expanded parts model for human attribute and action recogntion in still images, in: International Conference on Computer Vision, 2013, pp. 630–637.

[31] K. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Analy. Mach. Intel. 39 (9) (2010) 1582–1586.

[32] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, Multiple kernel for object deteciton, in: International Conference on Computer Vision, 1, 2009, pp. 92–100.

[33] S. Walk, N. Majer, K. Schindler, B. Schiele, New feature and insights for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 1, 2010, pp. 92–100.

[34] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 4, 2010. 3360–2167

[35] Q. Wang, J. Pang, L. Qin, S. Jiang, Q. Huang, Justifying the importance of color cues in object detection: a case study on pedestrian, in: Pacific-Rim Conference on Multimedia, 1, 2011, pp. 92–100.

[36] J. Weijer, C. Schmid, J. Verbeek, D. Larlus, Learning color names for real-world applications, IEEE Trans. Image Process. 18 (7) (2009) 1512–1524.

[37] C. Wojek, S. Walk, B. Schiele, Multi-cue onboard pedestrian detection, in: International Conference on Computer Vision, 1, 2009, pp. 92–100.

[38] J. Wu, J. Rehg, Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel, in: International Conference on Computer Vision, 2, 2009, pp. 630–637.

[39] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Conference on Commpute Vision and Pattern Recognition, volume 4, 20079, pp. 1794–1801.

[40] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, F.-F. Li, Human action recogntion by learning bases of action attributes and parts, in: International Conference on Computer Vision, 2, 2011, pp. 630–637.

[41] H. Zhang, A. Berg, M. Maire, J. Malik, Svm-knn: discriminative nearest neighbor classification for visual category recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 1, 2006, pp. 26–33.

[42] J. Zhang, K. Huang, Y. Yu, T. Tan, Boosted local structured hog-lbp for object location, in: IEEE Conference on Computer Vision and Pattern Recognition, 4, 2010, pp. 2126–2136.

[43] X. Zhen, L. Shao, X. Li, Action recogniton by spatio-temporal oriented energies, Inf. Sci. 281 (1) (2014) 295–309.

[44] C. Zhu, C. Bichot, L. Chen, Multi-scale color local binary patterns for visual object classes recognition, in: International Conference on Pattern Recognition, 2010, pp. 3065–3068.

[45] L. Zhu, Y. Chen, A. Yuille, W. Freeman, Latent hierarchical structural learning for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1062–1069.