Han *et al. IPSJ Transactions on Computer Vision and Applications* (2017) 9:4
DOI 10.1186/s41074-017-0017-4

IPSJ Transactions on Computer Vision and Applications

RESEARCH PAPER

Open Access

Co-occurrence context of the data-driven quantized local ternary patterns for visual recognition

Xian-Hua Han[1*], Yen-Wei Chen[2] and Gang Xu[2]

Abstract

In this paper, we describe a novel local descriptor of image texture representation for visual recognition. The image features based on micro-descriptors such as local binary patterns (LBP) and local ternary patterns (LTP) have been very successful in a number of applications including face recognition, object detection, and texture analysis. Instead of binary quantization in LBP, LTP thresholds the differential values between a focused pixel and its neighborhood pixels into three gray levels, which can be explained as the active status (i.e., positively activated, negatively activated, and not activated) of the neighborhood pixels compared to the focused pixel. However, regardless of the magnitude of the focused pixel, the thresholding strategy remains fixed, which would violate the principle of human perception. Therefore, in this study, we design LTP with a data-driven threshold according to Weber's law, a human perception principle; further, our approach incorporates the contexts of spatial and orientation co-occurrences (i.e., co-occurrence context) among adjacent Weber-based local ternary patterns (WLTPs, i.e., data-driven quantized LTPs) for texture representation. The explored WLTP is formulated by adaptively quantizing differential values between neighborhood pixels and the focused pixel as negative or positive stimuli if the normalized differential values are large; otherwise, the stimulus is set to 0. Our approach here is based on the fact that human perception of a distinguished pattern depends not only on the absolute intensity of the stimulus but also on the relative variance of the stimulus. By integrating co-occurrence context information, we further propose a rotation invariant co-occurrence WLTP (RICWLTP) approach to be more discriminant for image representation. In order to validate the efficiency of our proposed strategy, we apply this to three different visual recognition applications including two texture datasets and one food image dataset and prove the promising performance that can be achieved compared with the state-of-the-art approaches.

Keywords: Local ternary pattern, Data-driven quantization, Weber's Law, Co-occurrence statistic, Visual recognition

1 Introduction

Visual recognition has posed a significant challenge to the research community of computer vision due to interclass variability, e.g., illumination, pose, and inclusion. The rich context of an image makes the semantic understanding (object, pattern recognition) very difficult. Although the community has spent a lot of effort on image-categorization classification [1–4], which leads to very powerful intermediate representation of images such as the bag-of-feature (BoF) model [5, 6], it still has some space to improve the recognition performance in some specific recognition applications such as texture, food image, and biomedical data. The main idea of the popular BoF model is to quantize local invariant descriptors, for example, obtained by some interest point detector techniques [7, 8] and a description with SIFT [9], into a set of visual words [6]. The frequency vector of the visual words then represents the image, and an inverted file system is used for efficient comparison of such BOFs. Therein, the local descriptors for microstructure representation are often untouched, using standard, expert-provided ones such as the SIFT [9], PCA-Sift [10], and HOG [11], which result in a similar performance for image recognition. Further, since the learned codebook in BOF model has

*Correspondence: hanxh1216@gmail.com
[1]Graduate School of Science and Technology for Innovation, Yamaguchi University, Yamaguchi 753-8511, Japan
Full list of author information is available at the end of the article

of high diversity and irregularity, it is difficult to integrate the context information of co-occurrence for image representation.

On the other hand, microstructure simply using pixel neighborhoods as small as 3×3 pixels for local pattern representation has been proven to be possibly powerful for discriminating texture information, where the most popular one is a local binary pattern (LBP). LBP [12–14] characterizes each 3×3 local patch (microstructure) into a binary series by comparing the surrounding pixel intensity with the center one, which sets the bit of the surrounding pixel as 1 if its intensity is lager than the center one, otherwise 0. LBP is robust against uniform changes, and also easy for extension owing to the regularly decimated levels for all neighborhood pixels. However, due to the lack of spatial relationships among local textures, there have still serious disadvantages in the original LBP representation. Therein, an extension of the LBP called CoLBP [15, 16], has been proposed by considering the co-occurrence (spatial context) among adjacent LBPs, which proved promising performances on several classification applications [6]. In addition, by integrating orientation context, Nosaka et al. [17, 18] explored a rotation invariant co-occurrence among LBP, which is shown to be more discriminant on image classification compared to CoLBP. Qi et al. [19] proposed a pairwise rotation invariant CoLBP for texture representation and achieved the promising recognition performances in several visual recognitions.

Although LBP-based descriptors showed promising performances compared to other feature representations, it is obvious that LBP only thresholds the differential values between neighborhood pixels and the focused one to 0 or 1, which is very sensitive to noise existing in the processed image. Tan et al. extended LBP to local ternary pattern (LTP) [20], which considers the differential values between neighborhood pixels and the focused one as no or negative/positive stimulus, and successfully applied for face recognition under difficult lighting conditions. Given a pre-set positive threshold η, LTP [20] can obtain a series of ternary values for local pattern representation. However, regardless of the magnitude of the focused pixel, the pre-set threshold η remains fixed, which would violate the principle of human perception. Therefore, following the fact that human perception of a pattern depends not only on the absolute intensity of the stimulus but also on the relative variance of the stimuli, we propose to quantize the ratio between the neighborhood and the center pixels, which is equivalent to adaptively (data-driven) decide the quantization point according to the magnitude of the focus pixel. This proposed quantization strategy is inspired by Weber's law, a psychological law [21], which states that the noticeable change of a stimulus such as sound or lighting by a human being is a constant ratio of the original stimulus. When the stimulus has small magnitude, small change can be noticeable. Thus, we propose a data-driven quantized local ternary pattern (WLTP) based on Weber's law, which gives the activation status of the neighborhood pixels by data-driven thresholding the noticeable change according to the stimulus of the focus pixel: positively activated (magnitude: 1) if the ratio between the stimulus change and the focused one is larger than a constant η, negatively activated (magnitude: -1) if the change ratio is smaller than $-\eta$, not activated (magnitude: 0) otherwise. By incorporating a co-occurrence (here, i.e., spatial and orientation) context, we extend the proposed WLTP to a rotation invariant co-occurrence WLTP (RICWLTP), which is robust to image rotation and has the high descriptive ability among WLTP co-occurrences. Compared with the-state-of-the-art methods, our proposed strategy can achieve much better performance results for several visual recognitions including two texture datasets and one food image dataset.

2 Related work

In computer vision, local descriptors (i.e., features computed over limited spatial support) have been proved well adapted for matching and recognition tasks, as they are robust to partial visibility and clutter. The current popular one for local descriptors is SIFT feature, which is proposed in [9]. With the local SIFT descriptor, usually there are two types of algorithms for object recognition. One is to match the local points with SIFT features in two images, and the other one is to use the popular bag-of-feature model (BOF), which forms a frequency histogram of a predefined visual words for all sampled region features [6]. However, the extraction of SIFT descriptor itself is time-consuming, and then matching a large amount of them or quantizing them into a histogram in a large number of visual words, which are generally needed in BOF model for acceptable recognition of images, also exhausts a lot of time. On the other hand, some research works [22, 23] also have shown that it is possible to discriminate between texture patterns using pixel neighborhoods as small as 3×3 pixel region, which demonstrated that despite the global structure of the images, very good discrimination could be achieved by exploiting the distributions of such pixel neighborhoods. Therefore, exploiting such microstructures for representing images in the distributions of local descriptors has gained much attention and has led to state-of-the-art performances [24–28] for different classification problems in computer vision. The basic one of these approaches is local binary patterns (LBPs) introduce by Ojala et al. [22] as a means of summarizing local gray-level structures. As noted above, LBP is a simple yet efficient texture operator that labels pixels of an image by establishing thresholds for the neighborhood of each

pixel based on the value of the central pixel; the result is a binary number associated with each neighborhood pixel. For image representation, LBP index histogram is generally calculated as feature. Unfortunately, because of the packed single histogram, spatial relations among the LBPs are mostly discarded; thus, it results in the loss of global image information. Motivated by the co-occurrence concept such as in Co-HOG [29–31] and joint haar-like features [30], Nosaka et al. [17, 18] proposed to integrate context information (i.e., spatial and orientation) into the conventional LBP for achieving high descriptive capacity in image representation; the similar extension of LBP for integrating context information also can be seen in the recent work [19].

However, the LBP-based descriptors only set differential values between neighborhood pixels and the focused pixel to zero or one, and then it has high sensitivity to noise in the processed image that in turn degrades discriminant of image representation. Thus, as noted above, Tan et al. extended LBP to a local ternary pattern (LTP) approach [20], which considers differential values between neighborhood pixels and the focused pixel as either a negative/positive stimulus or no stimulus whatsoever. Next, the series of ternary values are combined into an LTP index. Given intensities $[I(\mathbf{x}), I(\mathbf{x}+\triangle\mathbf{x}_1), \cdots, I(\mathbf{x}+\triangle\mathbf{x}_l), \cdots, I(\mathbf{x}+\triangle\mathbf{x}_{L-1})]$ of focused pixel $\mathbf{x}$ and its L neighbors $\mathbf{x}+\triangle\mathbf{x}_l$ (displacement vector), LTP thresholds differential values $[I(\mathbf{x}+\triangle\mathbf{x})_0 - I(\mathbf{x}), \cdots, I(\mathbf{x}+\triangle\mathbf{x}_l) - I(\mathbf{x}), \cdots, I(\mathbf{x}+\triangle\mathbf{x}_{L-1}) - I(\mathbf{x})]$ as

$$G(I(\mathbf{x}+\triangle\mathbf{x}_l)-I(\mathbf{x})) = \begin{cases} 1 & I(\mathbf{x}+\triangle\mathbf{x}_l) - I(\mathbf{x}) > \eta \\ -1 & I(\mathbf{x}+\triangle\mathbf{x}_l) - I(\mathbf{x}) < -\eta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where η is the pre-set constant for thresholding the differential values. Then, the LTP index at $\mathbf{x}$ is defined as

$$LTP(\mathbf{x}) = \sum_{l=0}^{L-1} [G(I(\mathbf{x}+\triangle\mathbf{x}_l) - I(\mathbf{x})) + 1]\, 3^l \tag{2}$$

In the above equation, regardless of the magnitude of the focused pixel, the pre-set threshold η in LTP remains fixed, which violates the principle of human perception.

3 Data-driven quantized LTP and co-occurrence context

3.1 Weber's law

Ernst Heinrich Weber, an experimental psychologist in the nineteenth century, approached the study of the human response to a physical stimulus in a quantitative fashion and observed that the ratio of the increment threshold to the background intensity is a constant [32]. This observation shows that the just noticeable difference (JND) between two stimuli is proportional to the magnitude of the stimuli, which is well known as Weber's law and can be formulated as:

$$\frac{\triangle I}{I} = a \tag{3}$$

where $\triangle I$ denotes the increment threshold (just noticeable difference for discrimination) and I denotes the initial stimulus intensity; a is known as the *weber fraction*, which indicates that the proportion on the left hand of the equation remains constant in spite of the variance in I. Simply speaking, Weber's Law states that the size of the just noticeable difference is a constant proportion (a times) of the original stimulus value, which is the minimum amount that stimulus intensity must be changed in order to produce a noticeable variation in sensory experience.

3.2 Weber-based LTP: WLTP

According to Weber's law, the JND of a focused pixel in relation to its neighboring pixels is proportional to the intensity $I(\mathbf{x})$ of the focused pixel. Thus, we quantize different values $[I(\mathbf{x}+\triangle\mathbf{x}_0) - I(\mathbf{x}), \cdots, I(\mathbf{x}+\triangle\mathbf{x}_l) - I(\mathbf{x}), \cdots, I(\mathbf{x}+\triangle\mathbf{x}_{L-1}) - I(\mathbf{x})]$ between a focused pixel $\mathbf{x}$ and its $L-1$ neighbors $\{\mathbf{x}+\triangle\mathbf{x}_l\}(l = 0, 1, \cdots, L-1)$ to form a ternary series as follows:

$$G\left(\frac{I(\mathbf{x}+\triangle\mathbf{x}_l) - I(\mathbf{x})}{I(\mathbf{x})+\alpha}\right) = \begin{cases} 1 & \frac{I(\mathbf{x}+\triangle\mathbf{x}_l)-I(\mathbf{x})}{I(\mathbf{x})+\alpha} > \eta \\ -1 & \frac{I(\mathbf{x}+\triangle\mathbf{x}_l)-I(\mathbf{x})}{I(\mathbf{x})+\alpha} < -\eta \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where η is a predefined constant and α is a constant that avoids the case in which there is zero intensity (i.e., no stimulus); we always set α to one in our experimentation. Equation (3) adaptively quantizes differential values between the focus pixel and its neighboring pixels into a series of ternary codes; the WLTP index at $\mathbf{x}$ is defined as

$$\mathrm{WLTP}(\mathbf{x}) = \sum_{l=0}^{L-1} \left[G\left(\frac{I(\mathbf{x}+\triangle\mathbf{x}_l) - I(\mathbf{x})}{I(\mathbf{x})+\alpha}\right) + 1 \right] 3^l \tag{5}$$

In general LBP, neighborhood pixel number L is usually set to 8. Due to the detailed quantization (i.e., the ternary representation instead of binary), L is set to 4 to reduce computational cost. The lth displacement vector $\triangle\mathbf{x}_l$ is formulated as $\triangle\mathbf{x}_l = (\mathbf{r}\cos(\theta_l), \mathbf{r}\sin(\theta_l))$, where $\theta_l = \frac{360^o}{L} l$ and $\boldsymbol{r}$ is the scale parameter (i.e., the distance from the neighboring pixels to the focused pixel) of WLTP. As a result, as shown in Fig. 1, WLTPs have $N_P = 81$ (=3^L) possible patterns.

3.3 Integration of spatial and orientation contexts

LTP and WLTP discard information regarding spatial relationships between adjacent patterns; such information

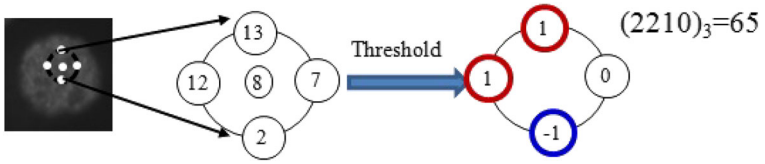

Fig. 1 The procedure of the proposed WLTP extraction

would be crucial to describing texture information for images. In this study, we first integrate the spatial context between the adjacent LTPs and WLTPs via co-occurrence information (called as CoLTP and CoWLTP, respectively). Without losing generalization, we describe the strategy of context integration for the proposed WLTP, while the same procedure can also be feasible by replacing WLTP with LTP. To obtain statistics of the co-occurrence (i.e., the spatial context) between two adjacent WLTPs, we consider the $N_P \times N_P$ auto-correlation matrix defined as:

$$H^{\varphi}_{p,q} = \sum_{\mathbf{x}\subset\mathbf{I}} \delta_{p,q}\left(\mathrm{WLTP}(\mathbf{x}), \mathrm{WLTP}(\mathbf{x}+\triangle\mathbf{x}_{\varphi})\right) \quad (6)$$

$$\delta_{p,q}(z_1, z_2) = \begin{cases} 1 & \text{if } z_1 = p \text{ and } z_2 = q \\ 0 & \text{otherwise} \end{cases}$$

where $p, q(=[0, 1, \cdots, N_P - 1])$ are possible pattern indexes for the two adjacent WLTPs and φ is the angle that determines the positional relations between the two WLTPs, which formulates displacement vector $\triangle\mathbf{x}_{\varphi} = (d\cos\varphi, d\sin\varphi)$ with interval d. Then, co-occurrence matrix dimension is 6561 $(= N_P \times N_P)$.

Further, because of the possible different imaging viewpoints, rotation invariant (i.e., orientation context) is generally an indispensable characteristic for texture image representation. Thus, we also integrate orientation context among adjacent WLTPs and propose a rotation invariant co-occurrence WLTPs that would contribute much higher descriptive capability for image representation. We first denote two pairs of WLTP patterns, $P^{\mathrm{WLTP}}_{\varphi=0} = [\mathrm{WLTP}(\mathbf{x}), \mathrm{WLTP}(\mathbf{x}+\triangle\mathbf{x}_{\varphi=0})]$ and $P^{\mathrm{WLTP}}_{\varphi} = [\mathrm{WLTP}^{\varphi}(\mathbf{x}), \mathrm{WLTP}^{\varphi}(\mathbf{x}+\triangle\mathbf{x}_{\varphi})]$, where $\mathrm{WLTP}(\mathbf{x})$ gives the 4-b clockwise ternary digits with the first digit in the right-horizontal direction ($\varphi = 0$), and $\mathrm{WLTP}(\mathbf{x}+\triangle\mathbf{x}_{\varphi=0})$ is the co-occurrence WLTP in the $\varphi = 0$ direction; $\mathrm{WLTP}^{\varphi}(\mathbf{x})$ and $\mathrm{WLTP}^{\varphi}(\mathbf{x}+\triangle\mathbf{x}_{\varphi})$ indicate the rotated entire WLTP pair with rotation angle φ. Thus, the rotation invariant statistics can be formulated if we assign the same index to P^{WLTP}_{φ} regardless of the different rotations designated by φ. Because we only used four neighbors of the focused pixel, only four rotation angles (i.e., $\varphi = 0°, 90°, 180°, 270°$) are available for computing rotation invariant statistics as shown in Fig. 2 (i.e., four equivalent WLTP pairs). According to the assigned labels for rotation invariant WLTP, the valid co-occurrence patterns can be reduced from 6561 to 2222. To efficiently calculate the rotation invariant co-occurrence of WLTP, we use mapping table **M** according to the algorithm shown in Table 1; the algorithm is to generate mapping table that converts a WLTP pair to an equivalent rotation index. In Table 1, $\mathrm{shift}((p)_3, l)$ means circle-shift l bits of the transformed ternary digits of p. With the calculated mapping table, statistics of the rotation invariant co-occurrence WLTP can be formulated as

$$H^{RI}_{M(\mathrm{Index})} = \sum_{\mathbf{x}\subset\mathbf{I}} \bigcup_{\varphi} \delta^{\mathrm{index}}\left[M(\mathrm{WLTP}^{\varphi}(\mathbf{x}), \mathrm{WLTP}^{\varphi}(\mathbf{x}+\triangle\mathbf{x}_{\varphi}))\right] \quad (7)$$

$$\delta^{\mathrm{index}}(z) = \begin{cases} 1 & \text{if } z = \mathrm{index} \\ 0 & \text{otherwise} \end{cases}$$

where $M(\mathrm{WLTP}^{\varphi}(\mathbf{x}), \mathrm{WLTP}^{\varphi}(\mathbf{x}+\triangle\mathbf{x}_{\varphi}))$ is the index of the rotation invariant co-occurrence between two WLTP pairs (RICWLTP). Finally, the statistics (i.e., histogram) of

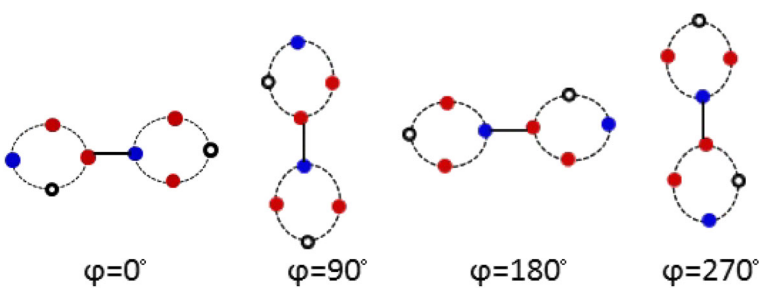

Fig. 2 The four equivalent WLTP pairs

Table 1 Calculation of mapping table between the WLTP pair and the rotation equivalent index

Mapping table generation algorithm

Input: Number of neighbor pixel L

Output: Mapping table $\mathbf{M}$ ($N_P \times N_P$ matrix)

Initialization: $Index$=1, $N_P = 3^L$, $\mathbf{M} \Leftarrow \{null\}^{N_P \times N_P}$

for: $p = 0, \cdots, N_P - 1$ and $q = 0, \cdots, N_P - 1$

if $M(p,q) = null$, **then**

$M(p,q) \Leftarrow Index, p' \Leftarrow shift((p)_3, 2)$,

$q' \Leftarrow shift((q)_3, 2), M(q', p') \Leftarrow Index$

$p' \Leftarrow shift((p)_3, 1)$,

$q' \Leftarrow shift((q)_3, 1), M(p', q') \Leftarrow Index$

$p' \Leftarrow shift((p)_3, 3)$,

$q' \Leftarrow shift((q)_3, 3), M(q', p') \Leftarrow Index$

$Index \Leftarrow Index + 1$

end if

end for

the RICWLTP can be used for discriminated representations of images. Similarly, if we formulate the Eq. (6) based on LTP instead of WLTP, the index of the rotation invariant co-occurrence between two LTP pairs (RICLTP) can be formed and then the statistics (i.e., histogram) of the RICLTP can be used for image representations.

3.4 Dimension analysis of LBP/LTP-based statistics

In general LBP, the used neighborhood pixel number L is usually set to 8, which results in 256 (2^8: 8 binary series) LBP indexes and the possible pattern value for any 3×3 local patch is ranged in $[0, 255]$. Thus, the LBP histogram for image representation has the dimension 256. The direct context integration considering the index combination of the adjacent LBP pairs would intuitively lead to a 256×256 co-occurrence matrix, where each element denotes the number of the LBP pairs with the same index combination in the image, and the reformed vector of the co-occurrence matrix is used as image representation with 65,536 (256^2) dimension. The high dimension of image representation results in high computational cost for the following classification procedure. Therefore, Nasaka et al. [17, 18] only adopted 4 ($L = 4$) neighborhood pixel number for producing 16 (2^4) LBP indexes, and then explored the co-occurrence statistics of LBP pairs (Co-LBP) and rotation invariant co-occurrence (RICLBP), which generated very compact LBP-based descriptors 256 ($(2^4)^2$) and 136 dimensional features, respectively. On the other hand, although Qi et al. [19] set the neighborhood pixel number L as 8, not all 256 LBP indexes are used for exploring co-occurrence of LBP pairs to avoid high-dimensional feature. In [19], the authors integrated the low-frequency appeared LBP indexes into one group for retaining a small number of LBP indexes (uniform LBP: ULBP, 59 patterns), and then further produced the rotation invariant uniform LBP (RIU-LBP, only 10 patterns) as the basic pair patterns. The co-occurrence matrix is formed by counting the number of the adjacent ULBP and RIU-LBP pairs with the combination indexes, which produces the matrix with a size of 59×10. The reformed vector as image representation is 590 in dimension and in [19], further integrating two different configurations of pair combination that generated 1180-dimensional features.

In (W)LTP-based descriptors, if the neighborhood pixel number L is set as 8, the number of LTP indexes is 6561 (3^8: 8 ternary series). In addition, the context integration considering the index combination of the adjacent (W)LTP pairs would generate an extremely large size of co-occurrence matrix (656×6561), which results in a 43,046,721-dimensional feature and is infeasible for post-processing. Thus, this study explores four neighborhood pixels for (W)LTP-based descriptors. For the simple histogram of the LTP and WLTP indexes (3^4: 4 ternary series), 81-dimensional image feature can be obtained, and the co-occurrence statistics (called CoLTP and CoWLTP) have the dimension 6561 (81^2). After considering the orientation-invariant property, our proposed RICLTP and RICWLTP produce a 2222-dimensional feature for image representation. Basically, the computational cost for extracting the LBP- and LTP/WLTP-based descriptors mainly depends on the pattern and pair-pattern index numbers, and thus, high-dimensional (W)LTP-based descriptors would lead to high computational cost. This is also a reason that this study applies only four neighborhood pixels, and it is also possible to explore different neighborhood structures for integrating more context information and multiscale patterns in our proposed RICWLTP, which is left to the future work.

4 Experiments

4.1 Data sets and methodology

We evaluated our proposed framework on two texture datasets and one food image dataset.

(i) Two texture datasets: the first one is Brodatz32 [33] that is a standard dataset for texture recognition, and the second one is KTH-TIPS 2a [34], a dataset for material categorization [35]. In Brodatz32 dataset, three additional images are generated by (i) rotating, (ii) scaling, and (iii) both rotating and scaling an original sample. We use the standard protocol [36], of randomly splitting the dataset into two halves for training and testing and report average performance over 10 random splits. KTH-TIPS 2a dataset contains 11 materials, e.g., cork, wool, linen, with images of four physical, planar samples for each material. The samples were photographed at nine scales, three poses,

and four different illumination conditions. It consists of 4395 images, most of which have the size of 200 × 200. All these variations on scale pose and illumination make it an extremely challenging dataset. For KTH-TIPS 2a texture dataset, we use the same evaluation protocol [36, 37] and report the average performance over four runs, where every time all images of one sample are taken for a test while the images of the remaining three samples are used for training.

(ii) Food dataset: Pittsburgh fast-food image dataset (PFID) [38], which is a collection of fast food images and videos from 13 chain restaurants, are acquired under lab and realistic settings. In our experiments, we focus on the set of 61 categories of specific food items (e.g., McDonald's Big Mac) with masked background. Each food category contains three different instances of the food (bought on different days from different branches of the restaurant chain) and six-viewpoint images (60° apart) of each food instance. We follow the experimental protocol in the published work [39] and perform threefold cross-validation for our experiments, using the 12 images from two instances for training and the 6 images from the third for testing. This procedure is repeated three times by using a different instance serving as the test set, and average performance is calculated as the results. The protocol ensures that no image of any given food item ever appears in both the training and test sets and guarantees that food items were acquired from different restaurants on different days. Two standard baseline algorithms, color histogram + SVM and bag of features with SIFT as the local descriptors + SVM, are shown for compared evaluation. In the baseline algorithm, the standard RGB color histogram with four quantization levels per color component is extracted to generate a $4^3 = 64$-dimensional representation for a food image, and a multi-class SVM is applied for classification. On the other hand, the BOF strategy combining a SVM classifier uses the histogram (low-order statistics) of visual words (representative words) with SIFT local descriptors. Recently, Yang et al. proposed Statistics of Pairwise Local Features (SPLF) [39] for food image representation, and the evaluation on PFID dataset showed the best performance on the state-

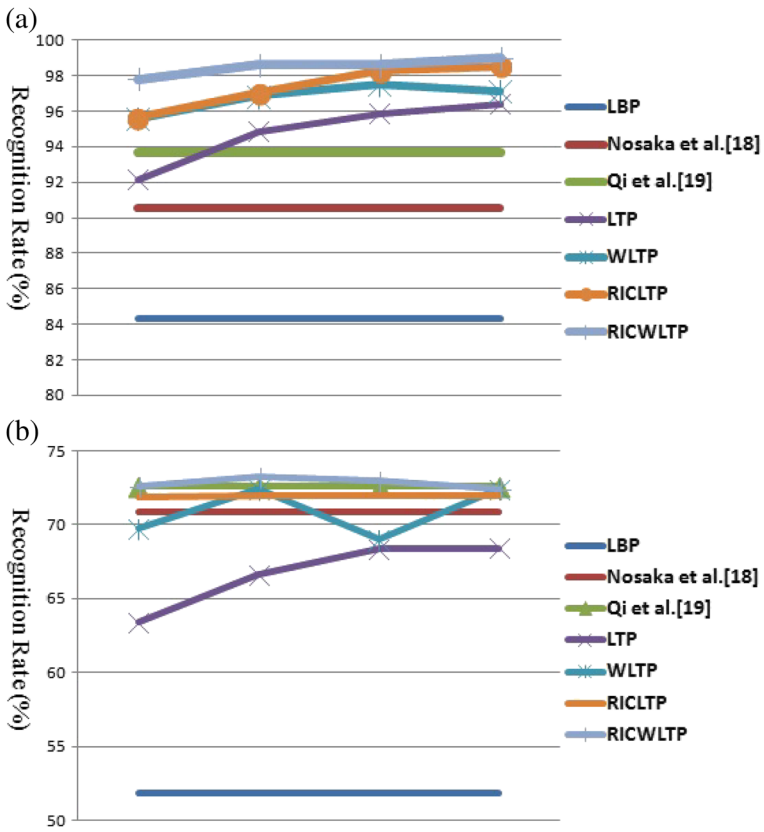

Fig. 3 Comparative results using our proposed framework and the conventional LBP-based descriptors by Nosaka et al. [17, 18] and Qi et al. [19]; *Horizontal axes* denotes the changed parameter with η introduced in Sections 2 and 3 for LTP, WLTP, RICLTP, and RICWLTP. Since there are no parameters in LBP-based descriptors, the recognition accuracies for LBP, the works by Nosaka et al. and Qi et al. remain the same in these graphs. **a** Bradatz dataset. **b** KTH-TIP 2a dataset

of-art methods, which are also as the compared results to our strategy.

4.2 Experimental results

We investigate the recognition performances using the statistics of our proposed WLTP, co-occurrence (i.e., spatial and orientation) context integration in WLTP, the corresponding versions without the using of Weber's law (not data-driven quantization, denote as LTP and RICLTP) and the conventional LBP and its extensions by Nosaka et al. [17, 18] and Qi et al. [19], which also incorporate the co-occurrence context into LBP. After obtaining the LBP/LTP-based features for image representation, we simply use linear support vector machine (SVM) due to its efficiency compared to the nonlinear one, for classification. In addition, we also pre-process the LBP/LTP-based histogram with the square root operation as the following:

$$\begin{aligned}\mathbf{P}'(\mathbf{I}) &= \left[p'_1(\mathbf{I}), p'_2(\mathbf{I}), \cdots, p'_L(\mathbf{I})\right] \\ &= \left[\sqrt{p_1(\mathbf{I})}, \sqrt{p_2(\mathbf{I})}, \cdots, \sqrt{p_L(\mathbf{I})}\right]\end{aligned} \tag{8}$$

where $\left[p_1(\mathbf{I}), p_2(\mathbf{I}), \cdots, p_L(\mathbf{I})\right] = \mathbf{P}(\mathbf{I})$ is the raw LBP/LTP/WLTP, context-integrated co-occurrence of LBP as the works by Nosaka et al. and Qi et al., and our proposed features, and L is the dimension of the focused features. $\mathbf{P}'(\mathbf{I})$ is the pre-processed or normalized histogram, which is used as the input of SVM for classification. With the above normalization, we can enhance some local patterns with low absolute frequency (values in histogram) in an image but large relative difference when comparing two images. Furthermore, a linear SVM with the processed features can also be explained as a nonlinear classification strategy using the raw LBP/LTP histogram, which is equivalent to using Hellinger Kernel in the raw feature space.

As introduced in Section 3, the (W)LTP and their context integration versions, RIC(W)LTP in the proposed framework is formulated by quantization procedure with a predefined (data-driven) threshold η. Figure 3 shows the comparative recognition performances using the features in works by Nosaka et al. and Qi et al., LBP, and the conventional LTP, the proposed WLTP, RICLTP, and

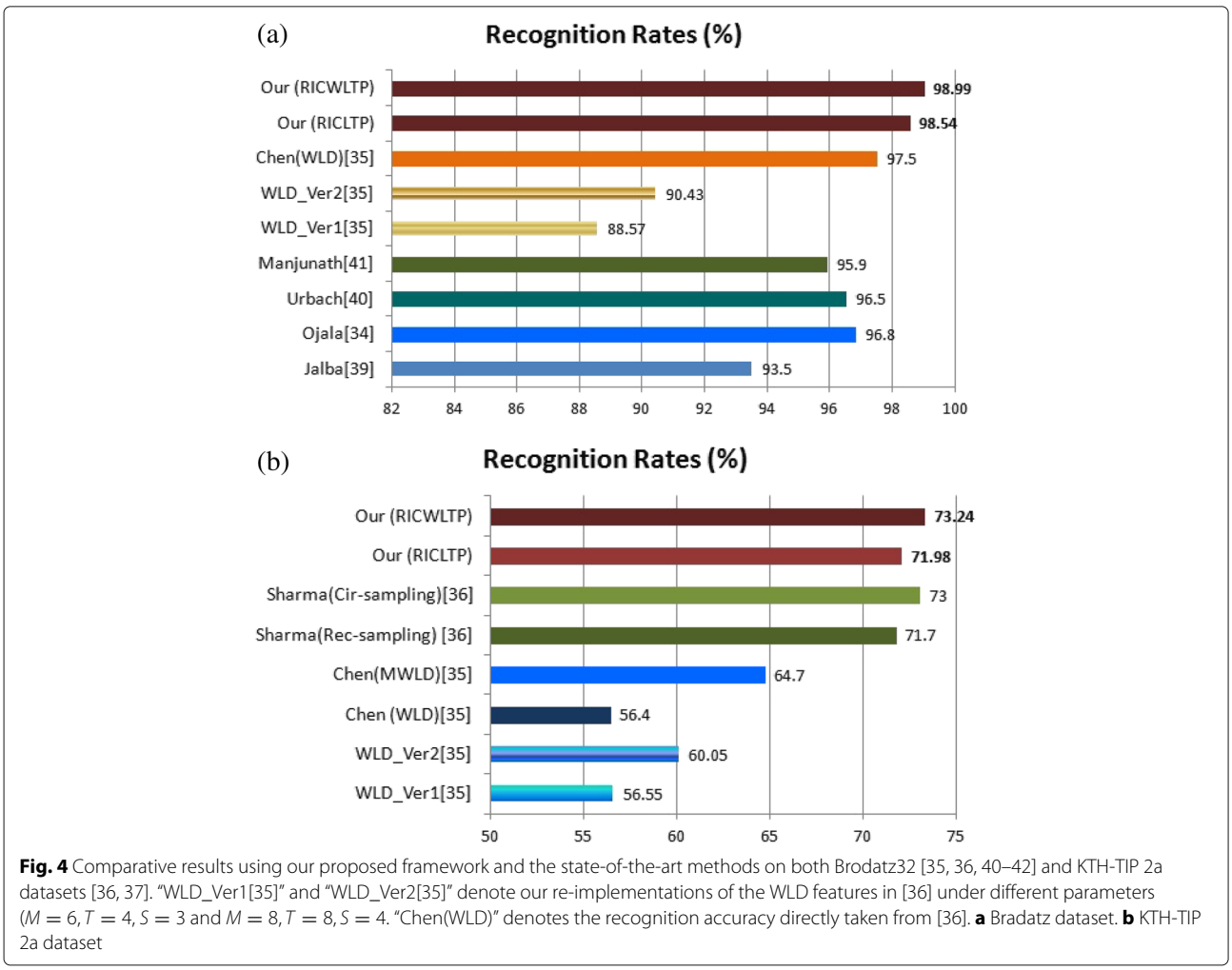

Fig. 4 Comparative results using our proposed framework and the state-of-the-art methods on both Brodatz32 [35, 36, 40–42] and KTH-TIP 2a datasets [36, 37]. "WLD_Ver1[35]" and "WLD_Ver2[35]" denote our re-implementations of the WLD features in [36] under different parameters ($M = 6, T = 4, S = 3$ and $M = 8, T = 8, S = 4$. "Chen(WLD)" denotes the recognition accuracy directly taken from [36]. **a** Bradatz dataset. **b** KTH-TIP 2a dataset

RICWLTP with different η on Brodatz32 and KTH-TIP 2a datasets; the LBP-based features (i.e., the works by Nosaka et al. and Qi et al., LBP) have no parameters. The parameters η in the LTP and RICLTP in our experiments are set as 1, 3, 5, and 7, and the ones (η) in the are set as 0.03, 0.05, 0.07, and 0.09. From Fig. 3, we observe that our proposed data-driven quantized versions (i.e., WLTP and RICWLTP) results in much better performance than that (i.e., LTP and RICLTP) with an absolute threshold, regardless of the magnitude of the focused pixel; the best recognition result was achieved by the proposed framework with data-driven quantization and co-occurrence context. Figure 4 gives the comparative recognition accuracies with our proposed frameworks and other state-of-the-art approaches [35–37, 40–42] on both Brodatz32 and KTH-TIP 2a datasets; the best results were achieved using our proposed approach with data-driven quantization and co-occurrence context. We also implemented the texture feature in [36] called the Weber local descriptor (WLD) for image representation and used the linear SVM with the normalized WLD histograms as Eq. (7) for classification. On both Brodatz32 and KTH-TIP 2a datasets, we extracted the histograms of WLD under different parameters ($M = 6, T = 4, S = 3$ and $M = 8, T = 8, S = 4$ as in [36], respectively, without block division for Brodatz32 but with nine block division for KTH-TIP 2a dataset) denoted as "WLD_Ver1" and "WLD_Ver2," respectively. Figure 4 shows that our proposed RIC(W)LTP methods can give much better performances on both texture datasets than WLDs implemented by [36] and us. All the above experiments were implemented using the linear SVM with the pre-processed LBP/LTP-based histograms by Eq. (8). As analyzed in [43], the linear SVM with the pre-processed image feature using root operator (Eq. (8)) is equivalent to the nonlinear SVM with Helinger kernel [43] on the raw image feature. This simple pre-processing combined with the linear SVM can achieve comparable classification performance with the nonlinear SVM and the raw image feature especially for histogram-based representation. For comparison, we also conducted the experiments using the nonlinear SVM with RBF kernel on both texture datasets. The compared recognition accuracies with RICLTP/RICWLTP and linear/nonlinear SVM are shown in Fig. 5a, b for Brodatz32 and KTH-TIP 2a datasets, respectively, which manifests that the comparable or a little better performances can be achieved by the nonlinear SVM than the linear one with the pre-processed RICLTP and RICWLTP descriptors. In the RBF kernel nonlinear SVM, there is a parameter γ, which is related to RBF kernel width and may greatly affect the classification performance. According to our experience [44, 45], we set $\frac{1}{\gamma}$ as the mean value of the pairwise Euclidean distances of all training samples' features in the nonlinear SVM experiments.

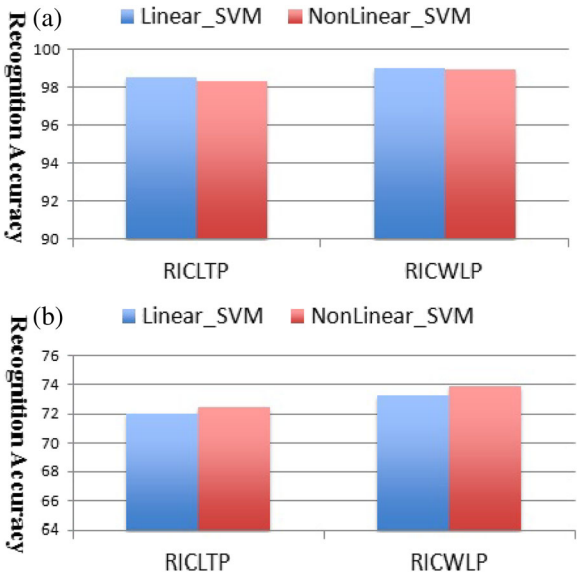

Fig. 5 Comparative results using linear SVM with pre-processed RICLTP and RICWLTP descriptors and the nonlinear SVM with RBF kernel on both texture datasets. **a** Bradatz dataset. **b** KTH-TIP 2a dataset

Next, we give the comparative results using our propose LTP-based descriptors and the conventional LBP-based descriptors (the features in the work by Nosaka et al. and Qi et al.) on PFID dataset in Fig. 6a that also shows the promising recognition performances by our strategy. Similar to the experiments for texture datasets, we also set the used parameter η in the LTP and RICLTP for food recognition experiments as 1, 3, 5, and 7 and η in the WLTP and RICWLTP as 0.03, 0.05, 0.07, and 0.09. For PFID database, the baseline evaluations by the color histogram and the BOF model using SIFT descriptor give the recognition rates 11.2 and 9.3%, respectively. In [39], the more promising performances on PFID database are achieved, which use statistics of pairwise local features. However, the proposed features need to firstly segment the different ingredients from the food image, which would cost more computational time. In our work, we apply the statistics of the simple data-driven quantization of the microstructures for food image representation and apply the very efficient linear SVM as classifier. The comparative results with the state-of-the-art methods are shown in Fig. 6b, where the best performance can be obtained by our proposed strategy. The more complex statistics of pairwise local features (denoted as SPLF) [39] can greatly improve the baseline evaluation (11.2, 9.3%) by conventional color histogram (denoted as CH) and the BOF model (denoted as BOF) to nearly three times (28.2%), which was the best performance on PFID in 2010. Our implementations of WLD histograms with different parameters [36] can also

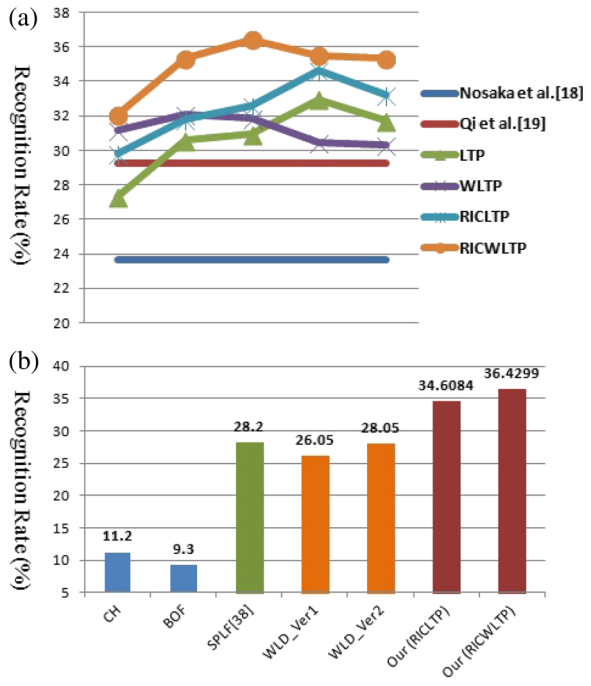

Fig. 6 Comparative results on PFID dataset **a** with the LBP based descriptors by Nosaka et al. [17, 18] and Qi et al. [19] and **b** with the state-of-the-art method [39]

give comparable performances with that by SPLF on PFID dataset, and our proposed strategy with data-driven quantization and co-occurrence context then can increase the recognition rate 28.2% with SPLF [39] to more than 36% about 8% improvement.

5 Conclusions

In this paper, we explored a robust representation strategy of texture images for visual recognition. The widely used local descriptor for texture analysis is local binary pattern, which characterizes each 3×3 local patch (microstructure) into a binary series by comparing the surrounding pixel intensity with the center one, and further has been extended to local ternary patterns via quantizing the difference values between the surrounding pixels and the center one into -1, 0, and 1, which can be explained as the active status (i.e., positively activated, negatively activated, and not activated). However, regardless of the magnitude of the focused pixel, the pre-set threshold η for quantization in the conventional LTP remains fixed, which would violate the principle of human perception. Motivated by the fact that human perception of a pattern depends not only on the absolute intensity of the stimulus but also on the relative variance of the stimuli, we proposed a novel local ternary pattern (LTP) with data-driven quantization according to Weber's law (called WLTP), which is equivalent to adaptively (data-driven) decide the quantization point according to the magnitude of the focus pixel. This proposed quantization strategy is inspired by Weber's law, a psychological law, which states that the noticeable change of a stimulus such as sound or lighting by a human being is a constant ratio of the original stimulus. Further, we incorporated the spatial and orientation context into WLTP and explored the rotation invariant co-occurrence among WLTP that has much higher descriptive capability than that of conventional LBP-based descriptors for image representation. Our experiments on two texture datasets and one food dataset confirmed that our proposed strategy greatly improved recognition performance as compared to the LBP-based descriptors and other state-of-the-art approaches.

Acknowledgements
This work was supported in part by the Grand-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports under Grant Nos. 15K00253, 15H01130, and 26330213; the Research Matching Fund for Private Universities from MEXT; and the Research Fund from Ritsumeikan University.

Authors' contributions
HXH carried out the method development, conducted experiments, participated in the sequence alignment, and drafted the manuscript. CYW and XG have revised the draft critically for important intellectual content and given the final approval of the version to be published. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

About the authors
Xian-Hua Han received a B.E. degree from ChongQing University, ChongQing, China, a M.E. degree from ShanDONG University, JiNan, China, a D.E. degree in 2005, from the University of Ryukyus, Okinawa, Japan. From Apr. 2007 to Mar. 2013, she was a post-doctoral fellow and an associate professor with the College of Information Science and Engineering, Ritsumeikan University, Japan. She is now a senior researcher at the Artificial Intelligence Researcher Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan. Her current research interests include image processing and analysis, feature extraction, machine learning, computer vision, and pattern recognition. She is a member of the IEEE, IEICE.
Yen-Wei Chen received a B.E.degree in 1985 from Kobe University, Kobe, Japan, a M.E. degree in 1987 and a D.E. degree in 1990 both from Osaka University, Osaka, Japan. From 1991 to 1994, he was a research fellow with the Institute of Laser Technology, Osaka. From Oct. 1994 to Mar. 2004, he was an associate professor and a professor with the Department of Electrical and Electronics Engineering, University of the Ryukyus, Okinawa, Japan. He is currently a professor with the College of Information Science and Engineering, Ritsumeikan University, Japan, and a professor with the Institute for Computational Science and Engineering, Ocean University of China, China. He is an Overseas Assessor of the Chinese Academy of Science and Technology, an associate Editor of the International Journal of Image and Graphics(IJIG), an Editorial Board member of the International Journal of Knowledge-Based Intelligent Engineering Systems and an Editorial Board member of the International Journal of Information. His research interests include intelligent signal and image processing, radiological imaging, and soft computing. He has published more than 100 research papers in these fields. Dr. Chen is a member of the IEEE, IEICE Japan and IEE(Japan).
Gang Xu received his Ph.D from Osaka University, in 1989. He was then a visiting researcher at ATR till 1990, when he returned to Osaka University as an assistant professor. In 1996, he became a tenured associate professor in Department of Computer Science, Ritsumeikan University. He was a visiting researcher at Robotics Lab, Harvard University in 1994, at Microsoft Research China in 1999, and at Motorola Australian Research Centre in 2000. Since 2001, he has been a full professor in Ritsumeikan University. In 2000, he founded 3D MEDiA Company Ltd. and served as CEO since then. His research interests

include image-based 3D metrology, 3D object recognition for industrial robots, and underlying techniques for pattern recognition in general. He authored and coauthored "Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach" (Kluwer Academic Publishers, 1996), "3D Vision" (Kyoritsu Shuppan in Japanese, 1999) and "3D CG from Photographs" (Kindai Kagakusha in Japanese, 2001).

Publisher's Note

Author details
[1]Graduate School of Science and Technology for Innovation, Yamaguchi University, Yamaguchi 753-8511, Japan. [2]Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan.

Received: 16 April 2016 Accepted: 1 March 2017
Published online: 14 March 2017

References

1. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. CVPR. IEEE, New York. pp. 2161–2168
2. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search. ECCV. Springer, Berlin
3. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2008) Lost in quantization: improving particular object retrieval in large scale image databases. CVPR. IEEE, New York
4. Han X-H, Chen Y-W, Ruan X (2012) Multilinear supervised neighborhood embedding of local descriptor tensor for scene/object recognition. IEEE Trans Image Process 21(3):1314–1326
5. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. ICCV. IEEE, New York. pp. 1470–1477
6. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. CVPR. IEEE, New York. pp. 2169–2178
7. Harris C, Stephens M (1998) A combined corner and edge detector. In: Proc. Alvey Vision Conference
8. Lowe D (1999) Object recognition from local scale-invariant features. Proc Int Conf Comput Vision. IEEE, New York. Vol. 2, pp. 1150-1157
9. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
10. Ke Y, Sukthankar R (2004) PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. Proc. IEEE Int Conf Comput Vision Pattern Recognit. IEEE, New York. pp. 506–513
11. Dalal N, Triggs B (2005) Histograms of Oriented Gradients for Human Detection. Proc. IEEE Int Conf Comput Vision Pattern Recognit. IEEE, New York. pp. 886-893
12. Wang XY, Han TX, Yan SC (2009) An HOG-LBP human detector with partial occlusion handling. ICCV. Springer, Berlin
13. Heikkila M, Pietikainen M (2006) A texture-based method for modeling the background and detecting moving objects. IEEE Trans Pattern Anal Mach Intell 28(4):657–662
14. Kertesz C (2011) Texture-based foreground detection. Int J Signal Process Image Process Pattern Recognit (IJSIP) 4(4):51–62
15. Louis W, Plataniotis KN (2011) Co-occurrence of local binary patterns features for frontal face detection in surveillance applications. EURASIP J Image Video Process
16. Nosaka R, Ohkawa Y, Fukui K (2011) Feature extraction based on co-occurrence of adjacent local binary patterns. In: The 5th Pacific-Rim Symposium on Image and Video Technology (PSIVT2011), LNCS. IEEE, New York Vol. 7088. pp 82–91
17. Nosaka R, Fukui K (2013) HEp-2 cell classification using rotation invariant co-occurrence among local binary patterns. Pattern Recogn 47:2428–2436
18. Nosaka R, Suryanto CH, Fukui K (2013) Rotation invariant co-occurrence among adjacent LBPs. Computer Vision: ACCV 2012, Workshops, Lecture Notes in Computer Science, Vol. 7728. Springer, Berlin. pp. 15–25
19. Qi XB, Xiao R, Zhang L, Guo J (2012) Pairwise rotation invariant co-occurrence local binary pattern. In: 12th European Conference on Computer Vision (ECCV2012). Springer, Berlin
20. Tan XY, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans Image Process 19(6):1635–1650
21. Shen JJ (2003) On the foundations of vision modeling I. Weber's law and Weberized TV (total variation) restoration. Physica D: Nonlinear Phenom 175(3/4):241–251
22. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 25:971–987
23. Varma M, Zissermann A (2003) Texture classification: are filter banks necessary? CVPRI. IEEE, New York. pp. 691–698
24. Xu Y, Yang X, Ling H, Ji H (2010) A new texture descriptor using multifractal analysis in multi-orientation wavelet pyramid. CVPR. IEEE, New York. pp. 161–168
25. Lazebnik S, Schmid C, Ponce J (2005) A sparse texture representation using local affine regions. PAMI 27:1265–1278
26. Zhang J, Marszalek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV 73:213–238
27. Varma M, Zisserman A (2005) A statistical approach to texture classification from single images. IJCV 62:61–81
28. Crosier M, Griffin LD (2010) Using basic image features for texture classification. IJCV 88:447–460
29. Watanabe T, Ito S, Yokoi K (2009) Co-occurrence histograms of oriented gradients for pedestrian detection. In: The 3rd IEEE Pacific-Rim Symposium on Image and Video Technology. IEEE, New York. pp 37–447
30. Kobayashi T, Otsu N (2008) Image feature extraction using gradient local auto-correlations. ECCV2008 5302:346–358
31. Mita T, Kaneko T, Stenger B, Hori O (2008) Discriminative feature co-occurrence selection for object detection. IEEE Trans Pattern Anal Mach Intell 30:1257–1268
32. Shen JJ, Jung Y-M (2006) Weberized Mumford–Shah model with Bose-Einstein photon noise. IJCV 53(3):331–358
33. Brodatz P (1966) Textures: a photographic album for artists and designers. Dover Publications, New York
34. Caputo B, Hayman E, Mallikarjuna P (2005) Class-specific material categorisation. In: the Proceedings of the 10th International Conference on Computer Vision. IEEE, New York
35. Ojala T, Valkealahti K, Oja E, Pietikainen M (2001) Texture discrimination with multidimensional distributions of signed gray level differences. Pattern Recogn 34(3):727–739
36. Chen J, Shan S, He C, Zhao G, Pietikainen M, Chen X, Gao W (2010) WLD: a robust local image descriptor. PAMI 32:1705–1720
37. Sharma G, Hussain S, Jurie F (2012) Local higher-order statistics (LHS) for texture categorization and facial analysis. ECCV. Springer, Berlin. Lecture Notes in Computer Science, vol 7578. pp.1–12
38. Chen M, Dhingra K, Wu W, Yang L, Sukthankar R, Yang J (2009) PFID: Pittsburgh Fast-Food Image Dataset. In: The 16th IEEE international conference on Image processing. IEEE, New York. pp 289–292
39. Yang SL, Chen M, Pomerleau D, Sukthankar R (2010) Food recognition using statistics of pairwise local features. Proc. IEEE Int Conf Comput Vision. IEEE, New York. pp. 2246–2256
40. Jalba AC, Wilkinson MHF, Roerdink JBTM (2004) Morphological hat-transform scale spaces and their use in pattern classification. Pattern Recogn 37(5):901–905
41. Urbach ER, Roerdink JBTM, Wilkinson MHF (2007) Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images. IEEE Trans Pattern Anal Mach Intell 29(2):272–285
42. Manjunath B, Ma W (1996) Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell 18(8):837–842
43. Leung T, Malik J (2013) Effects of image retrieval from image database using linear Kernel and Hellinger Kernel mapping of SVM. Int J Sci Eng Res 4(5):1184–1190
44. Han X-H, Chen Y-W (2011) Biomedical imaging modality classification using combined visual features and textual terms. Int J Biomed Imaging 2011
45. Han X-H, Chen Y-W, Ruan X (2012) Multilinear supervised neighborhood embedding of local descriptor tensor for scene/object recognition. IEEE Trans Image Process 21(3):1314–1326