Pattern Recognition 63 (2017) 542–550

Contents lists available at ScienceDirect

Pattern Recognition

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

HEp-2 staining pattern recognition using stacked fisher network for encoding weber local descriptor☆

Xian-Hua Han[a,*], Yen-Wei Chen[b,c]

[a] National Institute of Advanced Industrial Science and Technology, 2-3-26, Aomi, Koutou-ku, Tokyo 135-0064, Japan
[b] College of Information Science and Engineering, Ritsumeikan University
[c] College of Computer Science and Technology, Zhejiang University

ARTICLE INFO

MSC:
31-01
99-00

Keywords:
HEp-2 image representation
Fisher network
Stacked fisher framework
Weber's law
Weber local descriptor

ABSTRACT

This study addresses the recognition problem of the HEp-2 cell using indirect immunofluorescent (IIF) image analysis, which can indicate the presence of autoimmune diseases by finding antibodies in the patient serum. Generally, the method used for IIF analysis remains subjective, and depends too heavily on the experience and expertise of the physician. This study aims to explore an automatic HEp-2 cell recognition system, in which how to extract highly discriminate visual features plays a key role in this recognition application. In order to realize this purpose, our main efforts include: (1) a simple but robust local descriptor without any quantization for local patch representation; (2)A transformation of the difference values between the surrounding pixels and the center one to the perception degree, which is based on the fact that human perception for disguising a pattern depends not only on the absolute intensity of the stimulus but also on the relative variance of the stimulus; called as Weber local descriptor (WLD); (3) a data-driven coding strategy with a parametric probability process, and the extraction of not only low- but also high-order statistics for image representation called as Fisher vector; (4) the stacking of the Fisher network into multi-layer framework for more discriminate feature. Experiments using the open HEp-2 cell dataset released in the ICIP2013 contest validate that the proposed strategy can achieve a much better performance than the state-of-the-art approaches, and that the achieved recognition error rate is even very significantly below the observed intra-laboratory variability.

1. Introduction

Indirect immunofluorescence (IIF) is widely used as a diagnostic tool via image analysis; it can reveal the presence of autoimmune diseases by finding antibodies in the patient sera. Since it is effective for diagnosing autoimmune diseases [1], the demand for applying IIF image analysis in diagnostic tests is increasing. One research area involving IIF image analysis lies in the identification of the HEp-2 staining cell patterns using progressive techniques developed in the computer vision and machine learning fields. Several attempts to achieve the automatic recognition of HEp-2 staining patterns have been made. Perner et al. [2] proposed the extraction of texture and statistical features for cell image representation and then combined the extraction with a decision tree model for HEp-2 cell image classification. Soda et al. [3] investigated a multiple expert system (MES) in which an ensemble of classifiers was combined to label the patterns of single cells; however, research in the field of IIF image analysis is still in its early stages. There is still significant potential for improving the performance of HEp-2 staining cell recognition. Further, although several approaches have been proposed, they have usually been developed and tested on different private datasets under varying conditions, such as image acquisition according to different criteria and different staining patterns. Therefore, it is difficult to compare the effectiveness of these different approaches. In our study, we aim to achieve the automatic recognition of six HEp-2 staining patterns in an open HEp-2 dataset, which was recently released as part of the second HEp-2 cells classification contest at ICIP2013. There are a lot of works for exploring the recognition performances on this released HEp-2 cell dataset, and achieved promising results [4–9]. In the first HEp-2 cells classification contest at ICIP2012, it was shown that the LBP-based descriptor, rotation invariant co-occurrence LBP (RICLBP) for cell image representation, achieved promising HEp-2 cell classification performance [4,5]. In the second HEp-2 cells classification contest at ICIP2013, it was further shown that the combination of another extended LBP version, pairwise rotation invariant co-occurrence LBP (PRICoLBP) [10] and Bag-of-Features (BOF) [11] with a Sift descriptor

☆ National Institute of Advanced Industrial Science and Technology.

* Corresponding author.
E-mail addresses: hanxhua@fc.ritsumei.ac.jp (X.-H. Han).

http://dx.doi.org/10.1016/j.patcog.2016.09.025
Received 31 January 2016; Received in revised form 15 September 2016; Accepted 21 September 2016
Available online 22 September 2016

[12] achieved the best recognition results. LBP [13–15] characterizes each 3×3 local patch of an image into a binary series by comparing the surrounding pixel intensities with that of the center one, and setting the corresponding bit of a surrounding pixel as 1 if its intensity is larger than the center one; otherwise as 0, which is a procedure of binary quantization (coding) on the difference values between the surrounding and center pixels. Then, an 8-bit binary series can be obtained for each focused pixel (the center pixel) to forming a LBP index with a range [0–255]. The statistics (histogram, co-occurrence) of the LBP Index in an image can be extracted as image representation. However, the LBP index is computed only using the information if the surrounding pixel intensity is greater than that of the center one, and thus, the detail difference between them is thrown away with the binary coding procedure. Therefore, the very-rough quantization will lead to limited representation of local patch with the LBP. In addition, the LBP-based representation is commonly formed using a histogram, and hence, is restricted to the use of low-order statistics. In intuition, it would be more reliable to directly use the difference values between the surrounding pixels and the center one as the local descriptor, named as texton in this study. Therefore, Sharma et al. [16] proposed to directly use the local patch without any quantization as local descriptor and coded with Gaussian mixture model (GMM) [17] to extract high-order statistics for texture image representation, which has been proven to give promising recognition performance for several material datasets. Furthermore, Manivannan et al. [9] modeled multi-resolution local patterns with sparse coding and GMM for extracting discriminated features of HEp-2 cell images and manifested the impressive recognition performance on ICPR2014 contest HEp-2 cell dataset. On the other hand, the deep fisher network (DFN) with SIFT local descriptor has been proposed in [18] for general object recognition. Although DFN manifested outstanding classification performance for general image datasets with the popular local descriptor: SIFT [12] as the input of DFN, which is a gradient-weighted orientation histogram with rough quantization bins and thus leads to loss of the detail information, it would be not suitable for the fine-grained pattern recognition in HEp-2 cell dataset.

All the above mentioned feature extraction methods either use the popular local descriptor: SIFT, which leads to loss of the detail information, or the difference vector between the neighborhood and the center pixels, which complectly ignores the absolute magnitude of the center pixel and greatly affect the perception degree of the surrounding pixels by human. Therefore, in this study, we propose to adaptively normalize the difference values using the magnitude of the center pixel, and thus the normalized values can be considered as the perception degree of the surrounding pixels related to the center one. This insight of normalization is 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 stimulus, which is inspired by Weber's law. Weber's law, a psychological law [19], states that the noticeable change of a stimulus such as sound, lighting by human being is a constant ratio of the original stimulus. When the stimulus has a small magnitude, small changes can be noticeable. The normalized difference values between the surrounding pixels and the center one are formed as a vector for local patch representation, named as Weber local descriptor in this study. Several researchers also have used Weber's law in computer vision, where for example in [20], Weber's law is used to transform the raw image domain into an excitation domain, and then, directly concatenate the local descriptors in the excitation domain for image representation. This strategy first requires the normalization of the processed image into a uniform size, which therefore leads to very high-dimensional vectors for image representation. In this paper, a large amount of WLDs are also extracted from any cell image, and we model them using a general probability process in order to aggregate the large number of WLD. The used probability process is a Gaussian mixture model (GMM) [17,21]. Through modeling with GMM, we can achieve a data driven partition of the WLD space by learning parameters using training data, and aggregate the deviations to the learned average GMM parameters of the extracted WLD from an arbitrary image; the deviation vectors consist of not only low-order but also high-order statistics, which is also called as Fisher vector or Fisher network. In order to explore high-level features [18] for cell image representation, we further stack the Fisher network (called as stacked Fisher network: SFN) into multi-layer framework. Unlike the single layer Fisher network that directly encodes and summarizes all WLD of an input cell image as Fisher vector for image representation, the proposed SFN first aggregates the divinations (Fisher vectors: FVs) to the learned GMM parameters in densely sampled sub-region based on WLD, and then de-correlate and compresses these subregion-level FVs, and finally employs another FV for encoding the compressed subregion-level FVs.

Our primary contributions are four-fold: (1) a simple but robust local descriptor without any quantization for local patch representation; (2)A transformation of the difference values between the surrounding pixels and the center one to the perception degree, which is based on the fact that human perception for disguising a pattern depends not only on the absolute intensity of the stimulus but also on the relative variance of the stimulus; called as Weber local descriptor (WLD); (3) a data-driven coding strategy with a parametric probability process, and the extraction of not only low- but also high-order statistics for image representation called as Fisher vector; (4) the stacking of the Fisher network into multi-layer framework for more discriminate feature. Experimental results for the open HEp-2 cell dataset used at the ICIP2013 contest show that the variability of the recognition performance achieved by our proposed strategy is even significantly less than the observed intra-laboratory variability for both positive and intermediate intensity cell types. We also validate that the proposed strategy can achieve a much better performance than the state-of-the-art approaches.

The paper is organized as follows. Section 2 introduces the medical material for evaluating our proposed cell image representation. Our proposed stacked Fisher network is presented in Section 3 by first describing the basic idea for forming Weber local descriptor (WLD), and following the encoding procedure of WLD using Gaussian mixture model, also called Fisher network (FN), the stacking procedure into multi-layer FNs. Experimental results and conclusions are given in Sections 4 and 5, respectively.

2. Medical materials

In ANA tests, the HEp-2 substrates, in general, is applied, and both fluorescence intensity type and staining pattern are need to be distinguished, which is a challenging task affecting the reliability of IIF diagnosis. For fluorescent intensity, the Center for Disease Control and Prevention in Atlanta, Georgia (CDC) [22] established the guidelines, suggesting semi-quantitative scoring be performed independently by two physician IIF experts. The suggested score ranges from 0 to 4+ according to the intensity: negative (0), very subdued fluorescence (1+), defined pattern but diminished fluorescence (2+), less brilliant green (3+), and brilliant green or maximal fluorescence (4+). The values of the score are relative to the intensity of a negative and a positive control. The cell with positive intensity allows the physician to check the correctness of the preparation process, whereas that with negative intensity represents the auto-fluorescence level of the slide under examination. To reduce the variability of multiple readings, Rigon et al. [23] recently promoted to divide the fluorescence intensity into three classes, named negative, intermediate, and positive, by statistically analyzing the variability between several physicians' fluorescence intensity classification. In this study, we use the open ICIP2013 HEp-2 dataset to evaluate our proposed image representation strategy, which is a released dataset with only two intensity types: positive and intermediate, for evaluating the performances of different recognition techniques on staining pattern recognition.

Table 1
Cell image numbers for different staining patterns and different intensity types.

	Homogeneous	Speckled	Nucleolar	Centromere	NuMem	Golgi
Positive	1087	1457	934	1387	943	347
Intermediate	1407	1374	1664	1364	1265	377

The open ICIP2013 HEp-2 dataset includes intermediate and positive intensity types of HEp-2 cells; the purpose of the study involving this dataset is typically develop a means to recognize the staining pattern given the intensity types. Staining patterns primarily include the following six classes, with available image numbers for positive and intermediate intensity types shown in Table 1.

(1) Homogeneous: this pattern has the characterization with a diffuse staining of the interphase nuclei and staining of the chromatin of mitotic cells;

(2) Speckled: it has a granular nuclear staining of the interphase cell nuclei, which then consists of fine and coarse speckled patterns;

(3) Nucleolar: characterized by clustered large granules in the nucleoli of interphase cells that tend toward homogeneity, with fewer than six granules per cell;

(4) Centromere: characterized by several discrete speckles (~40–60) distributed throughout the interphase nuclei and characteristically found in the condensed nuclear chromatin during mitosis as a bar of closely associated speckles;

(5) Golgi: also called the Golgi apparatus, is one of the first organelles to be discovered and observed in detail. It is composed of stacks of membrane-bound structures known as cisternae;

(6) NuMem: NuMem is abbreviation from nuclear membrane, and has the characterization with a fluorescent ring around the cell nucleus, which are produced by anti-gp210 and anti-p62 antibodies.

In ICIP2013 HEp-2 dataset, there totally are over 10000 images, each showing a single cell, obtained from 83 training IIF images by cropping the bounding box of the cell. Example images for all six staining patterns from the positive and intermediate intensity types are shown in Fig. 1. Using the provided HEp-2 cell images and their corresponding patterns, we extract features that are effective for image representation, and learn a classifier (or a mapping function) using these extracted features of cell images and corresponding staining patterns. With the constructed classifier (the mapping function), the staining pattern can be automatically predicted given any HEp-2 cell image. In the next section, we describe our proposed feature extraction framework for cell image representation..

3. Stacked fisher network for encoding weber local descriptor

In this section, we describe our proposed framework for HEp-2 cell image representation. We first introduce Weber's law, which motivates the transformation of the changed magnitudes between the surrounding pixels and the center one for exploring local structures. Second, we describe our proposed data-driven model of the explored WLDs using GMM, and explore both low- and high-order statistics of the encoded WLDs for sub-region representation, also called the first layer Fisher network (FN). Finally, the second layer FN is introduced for modeling sub-region descriptor from the first layer FN.

3.1. Weber local descriptor

Recently, local descriptors of images (i.e. features computed over limited spatial support) have attracted much attention, and their statistical integration for image representation have been proved to be well-adapted for recognition tasks [13] since they are robust to partial visibility and clutter. The widely used technique for local descriptors in object recognition is SIFT feature, which is proposed

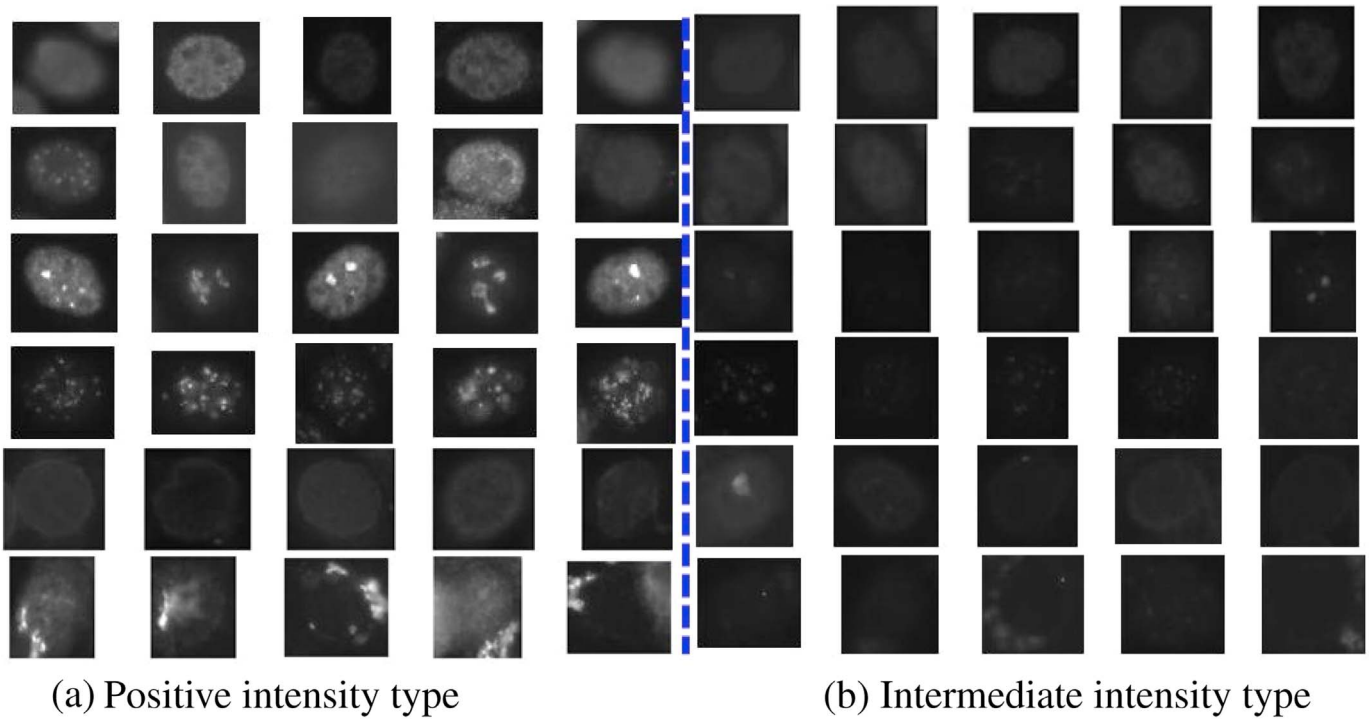


Fig. 1. Sample images for both positive and intermediate intensity types. Each row denotes the sample images of one staining pattern, and from up to down, the staining patterns are: Homogeneous, Speckled, Nucleolar, Centromere, NuMem and Golgi. (a) Positive intensity type; (b) Intermediate intensity type.

in [12]. SIFT descriptor firstly uniform-quantize the gradient directions into several bins (direction prototypes) in a limited spatial support for extracting compact representation. However, the uniform-quantization would throw away detail structure in the local patch, which leads to unrecoverable information loss, and thus possibly results in low performance especial in medical image analysis without distinctive structures. In order to aggregate the large amount of local descriptors from an image, the Bag-of-Feature model (BOF) is popularly used, which forms a frequency histogram of a predefined visual-words for all sampled region features [24–26]. However, the extraction of SIFT descriptor itself is time-consuming, and then quantizing them into a histogram for a large number of visual words, which are generally needed in BOF model for acceptable recognition of images, also requires a lot of computational time. In the other hand, some research works [27,28] also have shown that it is possible to discriminate between texture patterns using pixel neighborhoods as small as a 3×3 pixel region. Thus exploiting such the simple local patch for representing images in the distributions of local descriptors has gained much attention and has led to state-of-the-art performances [28–31] for different classification problems in computer vision. In addition, the simple binary quantization of the difference values between the surrounding pixels and center one in the used local patch, and its statistics (histogram and co-occurrence) have been applied to the HEp-2 cell staining pattern recognition, and shown the promising performances [32,10]. However, the LBP index is computed only using the information if the surrounding pixel intensity is greater than that of the center one, and thus, the detail difference between them is thrown away with the binary coding procedure. It is intuitive that it would be more reliable to directly use the difference values between the surrounding pixels and the center one as the local descriptor, named as Texton in this study. However, the direct use of the difference values complectly ignores the absolute magnitude of the center pixel, which greatly affect the perception degree of the surrounding pixels by human. Therefore, in this study, we propose to adaptively normalize the difference values using the magnitude of the center pixel, and thus the normalized values can be considered as the perception degree of the surrounding pixels related to the center one. This insight of normalization is 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 stimulus, which is inspired by Weber's law. Next, we will simply introduce Weber's Law, and then describe the extraction of our proposed local descriptor, names as Weber local descriptor (WLD).

Weber's Law: Ernst Heinrich Weber, an experimental psychologist in the 19th century, conducted the study about 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 [33], named as Weber's law. Weber's law is historically an important psychological law quantifying the perception of change in a given stimulus, which states that the just noticeable difference (JND) is in constant proportion to the original stimulus magnitude. This observation shows that the just-noticeable difference between two stimuli is proportional to the magnitude of the stimuli, and can be formulated as:

$$\frac{\triangle I}{I} = a \tag{1}$$

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 variance in I. Simply speaking, Weber's Law states that the size of the just noticeable difference (JND) 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.

Weber local description: As we mentioned in Section 1, it will be more reliable to represent a small local patch by directly suing the difference values between the surrounding pixels and the center one instead of binary quantization (forming LBP index). However, as Weber's law states that the just noticeable difference (JND) is different according to the magnitude of the original stimulus, and conclusively is a constant proportion of the original stimulus magnitude. Therefore, this study explores to transform the difference values between the surrounding pixels and the center one into perception degree as the local descriptors according Weber's law, called as Weber local descriptor (WLD).

Given a local patch in an image, let I_c denotes the stimulus magnitude at center pixel p_c, I_c^i ($i = 0, 1, \ldots, M-1$) is the intensity of the i^{th} neighbor of p_c, and M is the number of neighbors. The intuitive way is to directly use the difference values instead of binary quantization forming a vector as local descriptor $\mathbf{x} = [I_c^0 - I_c, I_c^1 - I_c, \ldots, I_c^{M-1} - I_c]^T$, called as Texton. However, this representation of local patch ignores the human perception principle, which would result in the similar local descriptor for the patches with very different structures; two examples are shown in Fig. 2 for the local Texton descriptors. Thus, this study proposes to normalize the difference values using the intensity of the center pixel (considered as the magnitude of the stimulus); and the normalized values can be considered as the perception degree of human compared the stimulus of the center pixel, which is formulated as the following:

$$x_c^w = \frac{(I_c^i - I_c)}{I_c + \alpha} \tag{2}$$

where α is a constant for avoiding zero division. Then, we use the normalized difference values to form a vector $\mathbf{x^w} = [x_0^w, x_1^w, \ldots, x_{M-1}^w]^T$, as patch representation, called Weber local descriptor (WLD). With the normalization using the center pixel stimulus, we can obtain the excitation magnitude of the surrounding pixels, and thus preserve more discriminating features than only using the absolute value of the pixel p_c. Intuitively, a positive value of x_c^w simulates the case in which the surroundings are lighter than the current pixel, whereas a negative value of x_c^w simulates the case in which the surroundings are darker than the current pixel. The two examples of WLD are also shown in Fig. 2..

Given an HEp-2 cell image, we work with all possible l×l neighborhoods (with l set to 3, 5,…, among others) for extracting the Weber local descriptor (WLD); i.e., $\mathbf{x^w} = [x_0^w, x_1^w, \ldots, x_{M-1}^w]^T$, where $M = l*l - 1$ is the surrounding pixel number. This WLD can capture the main salience change pattern, which would activate human perception, and therefor be much discriminant for image representation. Aggregating the large amount of WLD into a compact and discriminant vector for image representation has a crucial impact on the post-performance of image classification applications. Motivated the studies on image feature extraction in generic image classification [34] involving aggregate local descriptors extracted from the image into a histogram, such as BOF, LBP, we propose to exploit the distribution $p(\mathbf{x}^w|\mathbf{I})$ of the WLD space for a given image to represent the image. The following subsection describes our adaptive modeling approach: Fisher network, regarding the WLD.

3.2. The first layer FN

We denote the WLD space samples, which are randomly selected from training images, by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, where $\mathbf{x}_i \in \mathbf{R}^D$; further N is the sample number and D is the dimension of the WLD. Assuming that the WLD space samples have probability distribution as in a GMM, we can formulate

$$P(\mathbf{X}/\lambda) = \sum_{k=1}^{K_1} w_k N(\mathbf{X}/\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

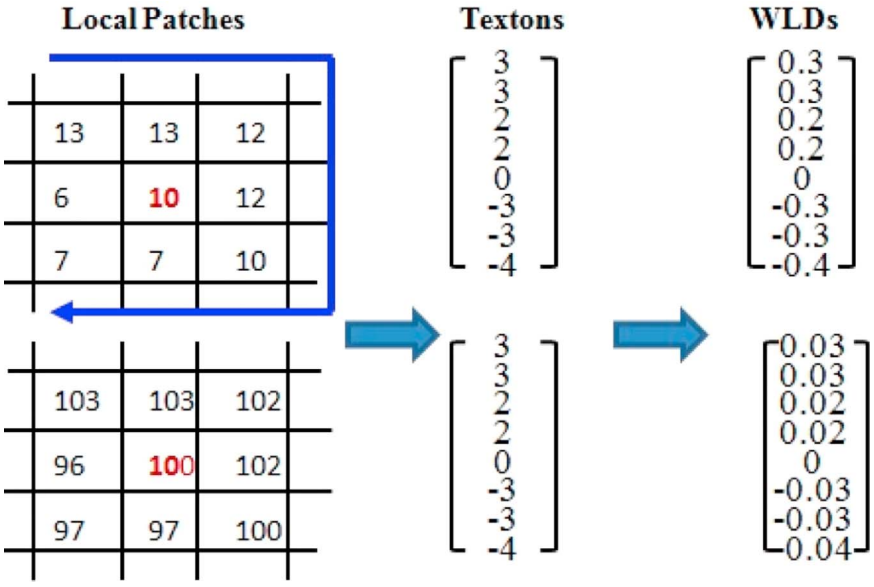


Fig. 2. The extracted Texton and WLD descriptors from two small patches. Texton gives the complete same descriptors for the two patches due to the same absolute changes between the surrounding pixels and the center one, while WLD manifests different representations according to the relative changes between the surrounding pixels and the center one which agrees to the perception principle of human. The arrow of the blue line denotes the taken order for forming descriptors.

where λ is the parameter for formulating the probability function in the GMM with K_1-components, denoted by $\lambda = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \ldots, K\}$. w_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ are the mixture weight, mean vector, and covariance matrix of Gaussian k, respectively, and $N(\mathbf{X}/\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian distribution with mean and covariance $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, respectively.

Given the training WLD samples, we can adaptively learn the prior parameters $\lambda = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \ldots, K_1\}$ of the GMM using an expectation maximization (EM) strategy [35] by maximizing the likelihood of the GMM of the training samples, which is equivalent to minimizing the (negative) log-likelihood as

$$L(\lambda) = -\sum_{i=1}^{N} \ln \sum_{k=1}^{K_1} w_k N(\mathbf{x}_i/\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4}$$

The EM strategy [35] iterates until it reaches a predefined iteration number or no (or minimal) change occurs in the above objective function. At that point, we identify the parameters $\lambda = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \ldots, K_1\}$ in the GMM that better fit the training texton samples.

The learned parameters (i.e., $\lambda = w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \ldots, K_1$) of the data-driven model (GMM) can fit into a WLD ensemble from a subregion of any HEp-2 cell image. The deviation statistics to the parameters are then the weight w_k (the 0^{th} order), mean $\boldsymbol{\mu}_k$ (the first order), and variance $\boldsymbol{\Sigma}_k$ (the second order); these can manifest the specific characteristics of the explored ensemble. These deviation statistics, also called high-order statistics or a Fisher vector, can be described via the Fisher kernel [36], and are given by the gradient of the log-likelihood of the data based on the learned model. It was proven in [36] that the utility of the Fisher kernel as the kernel machine in a discriminative classification model, which is inherently nonlinear, is equivalent to that of a linear kernel machine using the normalized deviation statistics as the feature vector. Therefore, the benefit of the explicit formulation for the Fisher vector is that a linear classifier can be used very efficiently.

For computational convenience, we assume that the weights are subject to the constraint: $\sum_{k=1}^{K_1} w_k = 1$, and using a D-dimensional micro-texton space, we assume that the covariance matrix is diagonal, denoted by $\boldsymbol{\sigma}_k = diag(\boldsymbol{\Sigma}_k)$. Given any texton sample $\mathbf{x}_t$ in the extracted texton set $\mathbf{X}$ (T texton) of a cell image, the occupancy probability for the k^{th} Gaussian component can be formulated as

$$\gamma_t(k) = \frac{w_k P(k/\mathbf{x}_t, \lambda)}{\sum_{k=1}^{K_1} w_k P(k/\mathbf{x}_t, \lambda)} \tag{5}$$

To explicitly avoid enforcing the constraints of weights w_k, we take a new relative parameter α_k to adopt soft-max formalism to define $w_k = \frac{\exp(\alpha_k)}{\sum_{j=1}^{K_1} exp(\alpha_j)}$. After re-parameterization using α_k and normalization with the Fisher information matrix $\mathbf{F}$ [37], the deviation for a WLD sample $\mathbf{x}_t$ from parameters $\lambda = \alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \ldots, K_1$ can be formulated as

$$\begin{aligned} \check{G}^{\mathbf{X}}_{\alpha_k} &= \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} [\gamma_t(k) - w_k], \quad \check{G}^{\mathbf{X}}_{\mu_k^d} = \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \left[\frac{x_t^d - \mu_k^d}{\sigma_k^d} \right] \check{G}^{\mathbf{X}}_{\sigma_k^d} \\ &= \frac{1}{\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \frac{1}{\sqrt{2}} \left[\frac{(x_t^d - \mu_k^d)^2}{(\sigma_k^d)^2} - 1 \right] \end{aligned} \tag{6}$$

where superscript d denotes the d^{th} dimension of the input vector $\mathbf{x}_t$, and k reflects the k^{th} Gaussian component in the learned model. Therefore, given a WLD ensemble $\mathbf{X}$, the aggregated deviation statistics from parameters α_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ can be considered as 0^{th}-order, first-order, and second-order statistics. The final feature for image or subregion representation is the concatenation of the deviation statistics with respect to all parameters, which can also be called as the Fisher vector (FV), and is of dimension $(2D + 1)K_1$.

To avoid dependence on the sample size, we normalize the resulting FV by the WLD sample size extracted from the given image or its subregion, i.e., $\check{\mathbf{G}}^{\mathbf{X}}_{\lambda} = \frac{1}{T}\check{\mathbf{G}}^{\mathbf{X}}_{\lambda}$.

3.3. The second layer FN

The single layer Fisher network aggregates (pools) all encoded WLDs in an image (pooling procedure), which only obtain the statistics (mid-level features) of the used WLDs (low-level features). In order to achieve much higher-level feature, we apply region-based pooling, which means achieving the statistics of the encoded WLD only from a sub-region by densely sampling image, and then normalize the pooled encoded vector for sub-region representation. Due to the high-dimension vector from the first Fisher network, we de-correlate them by principle component analysis (PCA) and compress by taking the PCs with accumulation contribution rate 90% for serving as the inputs of

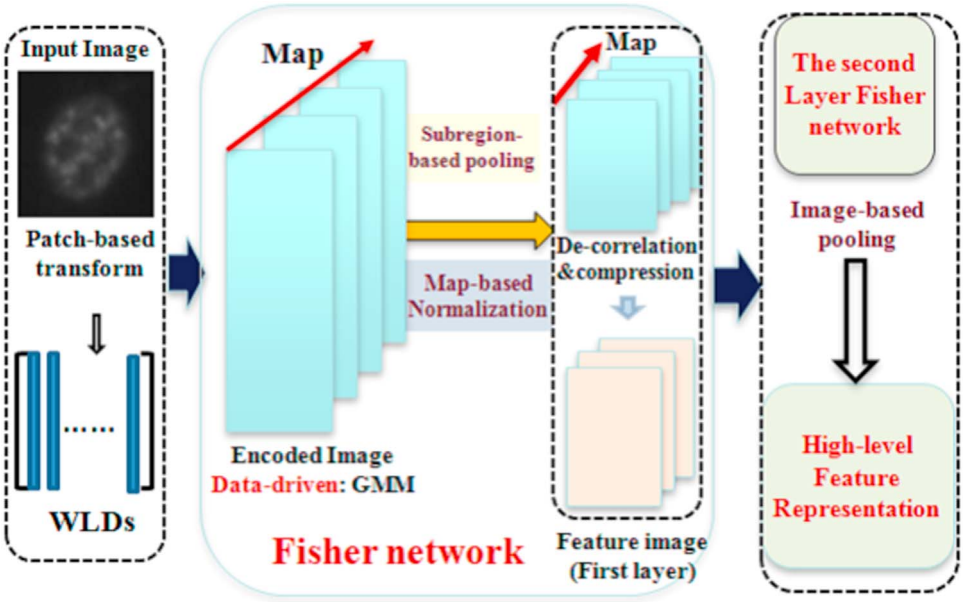
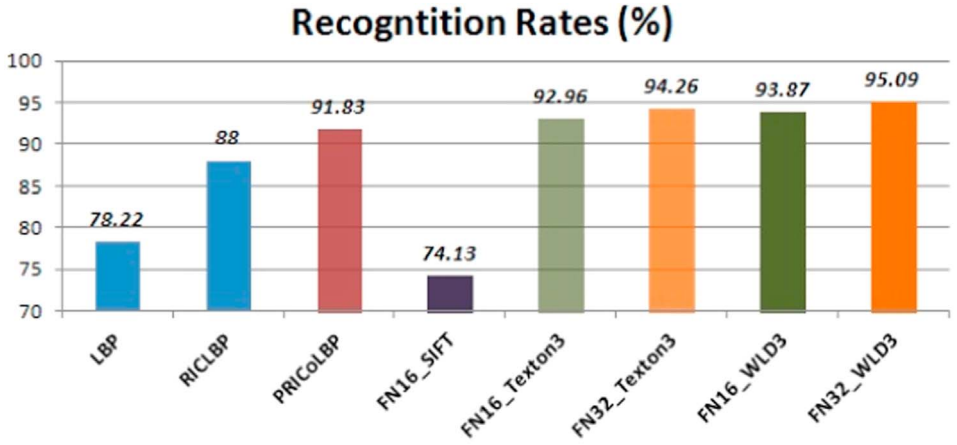
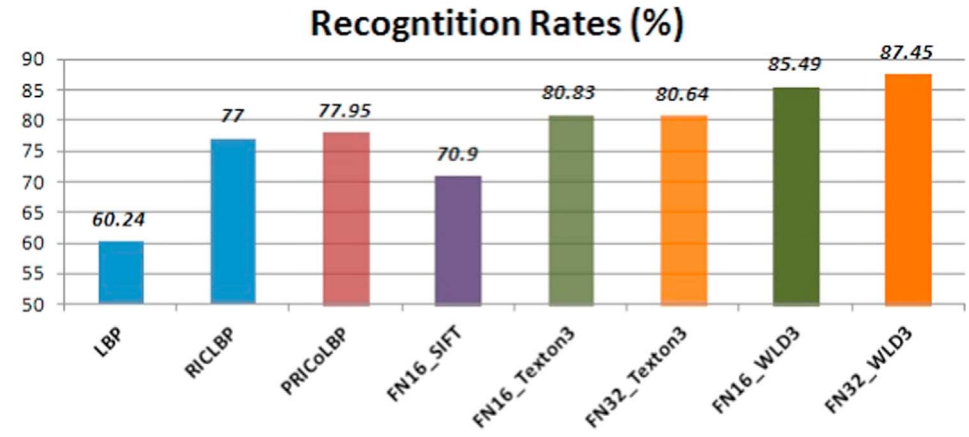


Fig. 3. The flowchart of our proposed stacked Fisher network.



(a) Compared result for positive intensity type



(a) Compared result for intermediate intensity type

Fig. 4. The compared recognition performance using one-layer FN-based features with different local descriptors (SIFT, Texton, WLD) and the LBP-based descriptors [10,32].

the 2nd FN layer. After learning a GMM with size of K_2, we apply another FN layer with these pre-processed subregion-based FVs and then extract the statistics of encoded subregion FVs over the entire image, which is prospected to have much higher-level vision than a single layer FN. Furthermore, we do element-wise operation on the FV feature with signed squared rooting following by l_2 normalization. The

flowchart of our proposed stacked FN is shown in Fig. 3..

4. Experimental results

Using the HEp-2 cell dataset, we validate the recognition performance of the two types of intensity (i.e., intermediate and positive) by applying the middle-level features using the single Fisher network and the conventional local binary pattern (LBP) and its extended versions, RICLBP [32] and the recently developed PRICoLBP [10], which have been proven to yield the promising recognition performance for the cell dataset [4]. The single FN (FN1: 16 and 32 Gaussian components, i.e. K_1=16, 32) with local Texton and our proposed WLD (with fixed patch size 3×3, and denoted as FN16_Texton3, FN32_Texton3, FN16_WLD3 and FN32_WLD3, respectively) and SIFT (denoted as FN16_SIFT) on the positive and intermediate intensity, separately. We conducted experiments with 10 fold cross validation on both positive and intermediate dataset, and the compared results are shown in Fig. 4. The linear SVM was used as the classifier because of its effectiveness as compared with other classifiers, such as K-nearest neighbor, and its efficiency as compared with a nonlinear SVM, which requires much more time to classify a sample. In addition, in order to achieve the acceptable recognition performance, we pre-processed LBP-based features using the square root operation for classification with a linear SVM:

$$\mathbf{f}' = [f_1, f_2, \ldots, f_L] = [\sqrt{f_1}, \sqrt{f_2}, \ldots, \sqrt{f_L}] \quad (7)$$

where $\mathbf{f} = [f_1, f_2, \ldots, f_L]$ is the raw histogram of LBP, RICLBP [32] or PRICoLBP [10] with dimension L, and $\mathbf{f}'$ is the normalized feature for linear SVM. This pre-processing of the LBP-based statistics for a linear classification model is equivalent to applying a nonlinear kernel, the Helinger kernel [38], for the raw histogram, and is expected to produce more promising results. In PRICoLBP, there exist several option parameters for radii (1 or 2) of neighbors and template numbers (1 or 2: two configurations, a or b); we conducted the recognition experiments on all the available PEICoLBPs with different parameters, and the best performance is achieved with the one of radii 2, template number 2 and configuration b, which is shown in Fig. 4. From Fig. 4, it is obvious that the FN for encoding local descriptor can give much better results than SIFT and LBP-based statistics, and the proposed WLD-based feature produces the best recognition performances. For HEp-2 cell classification, we use LIBSVM with the one-vs-rest strategy, and the cost parameter 'C' is fixed as 1 in all experiments..

The proposed WLD descriptor can be extracted from different patch sizes, and thus, the experiments were conducted for recognition performance evaluation with different sizes. Table 2 gives the experimental results using the one-layer FN with 3×3 and 5×5 patch sizes, which shows large size usually manifests better recognition performance. However larger patch size means high-dimensional local descriptors, and thus, leads to high computational time for encoding. We also tried larger size patch than 5×5, no improvement of recognition performance is observed for positive intensity type, and only a little improvement is shown in intermediate intensity type. In addition, Table 3 gives the compared performance with one and two- layer FN. From Table 3, it can be seen that the recognition rates can be increased using two-layer FNs than only one under different experimental conditions. The confusion matrix of HEp-2 cell recognition using two-layer FN-based features are shown in Table 4 on both positive and intermediate intensity types. Fig. 5 manifests the recognition performances with different Gaussian components (16/32) of the first and the second layers, where the first number is the component number of the first layer while the second is that of the second layer..

Finally, we compare the experimental results using our proposed stacked FN with WLD and the used features in [39], the proposed image representations in [9], and the compared results are shown in Table 5. It should be noted that our proposed stacked FN features were

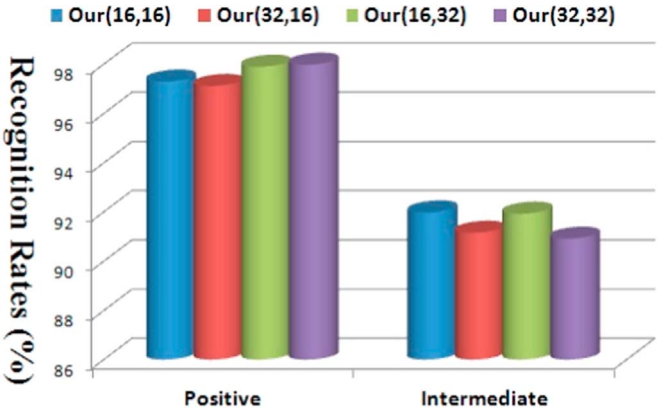


Fig. 5. The compared recognition performance with different Gaussian components.

Table 2
The compared recognition performances using the one-layer FN-based features with Texton and WLD on different experimental conditions.

Local descriptor	Recognition rates (%)			
	Experimental conditions			
	Patch size 3×3		Patch size 5×5	
	GMM numbers			
	16	32	16	32
(a) For positive intensity type				
Texton	92.96	94.26	95.3	*96.11*
WLD	93.87	95.09	95.58	**96.4**
(b) For intermediate intensity type				
Texton	80.83	80.64	83.88	87.62
WLD	84.06	86.08	*88.34*	**89.19**

Table 3
The compared recognition performances using the one-layer and two-layer FN-based features.

Local descriptor	Recognition rates (%)			
	Experimental conditions			
	Patch size 3×3		Patch size 5×5	
	GMM numbers of the first FN			
	16		16	
(a) For positive intensity type				
Texton	92.96		95.3	
WLD	93.87		95.58	
	GMM number of the second FN			
	16	32	16	32
Texton	96.28	96.55	*96.67*	*96.68*
WLD	95.23	96.8	**97.3**	**97.9**
(b) CFor intermediate intensity type				
Texton	80.83		83.88	
WLD	84.06		88.34	
	GMM number of the second FN			
	16	32	16	32
Texton	85.92	85.75	*89.72*	*89.61*
WLD	89.98	91.19	**91.98**	**91.93**

Table 4
The confusion matrix for classifying different HEp-2 cell patterns in the positive and intermediate type dataset using two-layer FN-based features (the Gaussian number of the first- and second-layer are: 16 and 32, respectively) with patch size 5×5.

%	Homogeneous	Speckled	Nucleolar	Centromere	NuMem	Golgi
(a) Confusion matrix for positive intensity type						
Homogeneous	**96.87**	2.68	0	0	0.45	0
Speckled	2.34	**97.38**	0.06	0	0.06	0.15
Nucleolar	0	0.42	**98.42**	0.96	0.10	0.10
Centromere	0.07	0.63	0.57	**98.45**	0	0.29
NuMem	0.74	0.32	0.11	0	**98.84**	0
Golgi	0.30	0.56	0.36	0.96	0.53	**97.28**
(b) Confusion matrix for intermediate intensity type						
Homogeneous	**95.49**	1.76	1.49	0	1.11	0.15
Speckled	2.66	**90.46**	4.41	1.52	0.95	0
Nucleolar	1.68	5.35	**88.52**	0.48	2.05	1.92
Centromere	0.07	1.97	1.40	**96.34**	0.15	0.07
NuMem	1.27	1.81	3.09	0.15	**93.12**	0.56
Golgi	3.73	0.24	11.94	0.53	4.57	**78.9**

Table 5
The compared results using our two-layer FN-based features and the used features in [39], the proposed image representation in [9]. The percentages in red color mean the recognition rates using the stacked FNs, and the ones in blue color means the best results with the image representation proposed in [41].

%	HOG	LBP	GLRL	SGLD	Laws	rSIFT[9]	MP[9]	Our
(a) For positive intensity type								
SVM_Linear	81.89	79.15	77.23	84.37	94.68	91.9	95.29	**97.90**
SVM_RBF	86.30	84.21	84.91	90.81	*97.90*	*	*	*
(b) For intermediate intensity type								
SVM_Linear	67.83	62.96	39.33	49.75	81.06	78	86.91	**91.93**
SVM_RBF	73.19	71.86	49.96	58.45	*90.49*	*	*	*

only evaluated with linear SVM (denoted as SVM_Linear) classifier. In [39], the used features mainly consist of HOG, LBP and three texture features: GLRL [40], Laws [41], SGLD [42], and the Laws texture features proposed in [41] gave the best HEp-2 cell recognition performance with nonlinear SVM (the SVM with RBF kernel, denoted as SVM_RBF) in [39]. Since [9] only gave the final recognition performance on the HEp-2 cell dataset by combining several local features such as the combined rSIFT and MP, the combined rSIFT and rotated MP, and so on. In order to give fair comparison, we implemented the feature representation in [9] by combining sparse coding and two local features: multi-resolution local pattern (denoted as MP) and Root-SIFT (denoted as rSIFT), and then conducted experiments using the same conditions as in our proposed framework. The multi-resolution local patterns in our implementation takes three-scales (radii: 1, 2, 3) and 8 sampling regions in each scale, and thus form a 24-dimensional MP vectors as local features. From Table 5, we can see that the proposed stacked FN features even with the linear SVM can achieve better or comparable recognition performance than the best results in [39] with nonlinear SVM, and also much better than the feature representation (our implementation) proposed in [9].

5. Conclusions

In this paper, we explored a robust local descriptor (called as WLD) inspired by Weber's law and its high-order statistics for HEp-2 cell image representation. Via modeling the WLDs with a parametric probability process, we can extract middle-level features for image sub-region representation, and further stack the above procedure into deep framework for high-level feature extraction, which is called as stacked Fisher network (SFN). Experiments on the HEp-2 cell dataset from ICIP2013 validated that our proposed strategy achieves the best recognition performance as compared with existing state-of-the-art approaches.

Acknowledgements

This research was supported in part by the Grant-in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 15K00253, 16H01436 and No. 15H01130, in part by the MEXT Support Program for the Strategic Research Foundation at Private Universities (2013-2017), and in part by the New Energy and Industrial Technology Development Organization (NEDO), the R-GIRO Research Fund from Ritsumeikan University, Japan, and in part by the Recruitment Program of Global Experts (HAIOU Program) from Zhejiang Province, China.

References

[1] M. Mahler, S. Pierangeli, P.-L. Meroni, M.J. Fritzler, Autoantibodies in systemic autoimmune diseases, Journal of Immunology Research 2014.
[2] P. Perner, H. Perner, B. Muller, Mining knowledge for hep-2 cell image classification, J. Artif. Intell. Med. 26 (2002) 161–173.
[3] P. Soda, G. Iannello, Aggregation of classifiers for staining pattern recognition in antinuclear autoantibodies analysis, IEEE Trans. Inf. Technol. Biomed. 13 (3) (2009) 322–329.
[4] P. Foggia, G. Percannella, P. Soda, M. Vento, Benchmarking hep-2 cells classification methods, IEEE Trans. Med. Imaging 32 (10) (2013) 1878–1889.
[5] P. Foggia, G. Percannella, A. Saggese, M. Vento, Pattern recognition in stained hep-2 cells: where are we now?, Pattern Recognit. 47 (2014) 2305–2314.
[6] A. Wiliem, C. Sanderson, Y. Wong, P. Hobson, R.F. Minchin, B.C. Lovell, Automatic classification of human epithelial type 2 cell indirect immunofluorescence images using cell pyramid matching, Pattern Recognit. 7 (2014) 2315–2324.
[7] L. Liu, L. Wang, Hep-2 cell image classification with multiple linear descriptors, Pattern Recognit. 7 (2014) 2400–2408.
[8] S. Manivannan, W. Li, S. Akbar, R. Wang, Hep-2 cell classification using multi-resolution local patterns and ensemble svms, I3A 1st workshop on Pattern Recognition Techniques for Indirect Immunoflurescence Images on ICPR 2014 (2014) 37–40.

[9] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, S.J. McKenna, An automated pattern recognition system for classifying indirect immunofluorescence images of hep-2 cells and specimens, Pattern Recognit. (2016) 12–26.
[10] X. Qi, R. Xiao, C. Li, J. Guo, X. Tang, Pairwise rotation invariant co-occurrence local binary pattern, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2199–2213.
[11] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, CVPR (2006) 2169–2178.
[12] D. Lowe, Distinctive image features from scale-invariant keypoint, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[13] D. He, L. Wang, Texture unit, texture spectrum, and texture analysis, IEEE Trans. Geosci. Remote Sens. 28 (1990) 509–512.
[14] X. Wang, T.X. Han, S. Yan, An hog-lbp human detector with partial occlusion handling, ICCV.
[15] G. Zhao, M. Pietikainen, Local binary pattern descriptors for dynamic texture recognition, Pattern Recognit. (2006) 211–214.
[16] G. Sharma, S. ul Hussain, F. Jurie, Local higher-order statistics (lhs) for texture categorization and facial analysis, European Conference on Computer Vision (ECCV2012) (2016) 1–12.
[17] I. Dinov, Expectation maximization and mixture modeling tutorial, California Digital Library, Statistics Online Computational Resource, Paper EM_MM, ⟨http://repositories.cdlib.org/socr/EM_MM⟩.
[18] L. Simonyan, A. Vedaldi, A. Zisserman, Deep fisher networks for large-scale image classification, Adv. Neural Inf. Process. Syst. (NIPS2013) (2012) 163–171.
[19] J.J. Shen, On the foundations of vision modeling i. weber's law and weberized tv (total variation) restoration, Phys. D: Nonlinear Phenom. (3/4) (2003) 241–251.
[20] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, W. Gao, Wld: a robust local image descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 9 (2010) 1705–1720.
[21] B. Christopher, Pattern Recognition and Machine Learning, Springer, New York.
[22] C. for Disease Control, Quality assurance for the indirect immunofluorescence test for autoantibodies to nuclear antigen (if-ana): approved guideline, NCCLS I/LA2-A 16(11).
[23] A. Rigon, P. Soda, D. Zennaro, G. Iannello, A. Afeltra, Indirect immunofluorescence in autoimmune diseases: assessment of digital images for diagnostic purpose, Cytom. B (Clin. Cytom.) 72 (3) (2007) 472–477.
[24] J. Hervé, M. Douze, C. Schmid, Improving bag-of-features for large scale image search, Int. J. Comput. Vis. 3 (2010) 316–336.
[25] L. Liu, L. Wang, X. Liu, in: defense of soft-assignment coding, International Conference of Computer Vision (2011) 2486–2493.
[26] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, CVPR2009.
[27] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, PAMI (2002) 971–987.
[28] Y. Xu, X. Yang, H. Ling, H. Ji, A new texture descriptor using multifractal analysis in multi-orientation wavelet pyramid, CVPR.
[29] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using local affine regions, PAMI (2005) 1265–1278.
[30] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, IJCV (2007) 213–238.
[31] M. Varma, A. Zisserman, A statistical approach to texture classification from single images, IJCV (2005) 61–81.
[32] R. Nosaka, K. Fukui, Hep-2 cell classification using rotation invariant co-occurrence among local binary patterns, Pattern Recognit. 27 (7) (2013) 2428–2436.
[33] J.J. Shen, Y.-M. Jung, Weberized mumford–shah model with bose-einstein photon noise, IJCV, vol. 3, 2006, pp. 331–358.
[34] L. Bo, X. Ren, D. Fox, Multipath sparse coding using hierarchical matching pursuit, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 660–667.
[35] L. Xu, M. Jordan, On convergence properties of the em algorithm for gaussian mixtures, Neural Comput. 1 (1996) 129–151.
[36] U. Dick, K. Kersting, Fisher kernels for relational data, ECML 2006: 17th European Conference on Machine Learning (2006) 114–125.
[37] F. Perronnin, J. Sanchez, T. Mensink, Improving the fisher kernel for large-scale image classification, European Conference on Computer Vision (ECCV2010) (2010) 143–156.
[38] T. Leung, J. Malik, Effects of image retrieval from image database using linear kernel and hellinger kernel mapping of SVM, International Journal of Scientific and Engineering Research, 4.
[39] P. Agrawal, M. Vatsa, R. Singh, Hep-2 cell image classification: A comparative analysis, Machine Learning in Medical Imaging, Lecture Notes in Computer Science, 2013, pp. 195–202.
[40] R. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man and Cybernetics, 6, 1973, pp. 610–621.
[41] K. Laws, Textured image segmentation, Technical report, USC.
[42] X. Tang, Texture information in run-length matrices, IEEE Transactions on Image Processing, 11, 1998, pp. 1602–1609.

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 April 2007 to March 2013, she was a post-doctoral fellow and an associate professor with College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. She is currently a senior researcher with the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, 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 October 1994 to March 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).