DERF: Distinctive Efficient Robust Features From the Biological Modeling of the P Ganglion Cells

Dawei Weng, Yunhong Wang, *Member, IEEE*, Mingming Gong, Dacheng Tao, *Fellow, IEEE*, Hui Wei, and Di Huang, *Member, IEEE*

***Abstract*— Studies in neuroscience and biological vision have shown that the human retina has strong computational power, and its information representation supports vision tasks on both ventral and dorsal pathways. In this paper, a new local image descriptor, termed distinctive efficient robust features (DERF), is derived by modeling the response and distribution properties of the parvocellular-projecting ganglion cells in the primate retina. DERF features exponential scale distribution, exponential grid structure, and circularly symmetric function difference of Gaussian (DoG) used as a convolution kernel, all of which are consistent with the characteristics of the ganglion cell array found in neurophysiology, anatomy, and biophysics. In addition, a new explanation for local descriptor design is presented from the perspective of wavelet tight frames. DoG is naturally a wavelet, and the structure of the grid points array in our descriptor is closely related to the spatial sampling of wavelets. The DoG wavelet itself forms a frame, and when we modulate the parameters of our descriptor to make the frame tighter, the performance of the DERF descriptor improves accordingly. This is verified by designing a tight frame DoG, which leads to much better performance. Extensive experiments conducted in the image matching task on the multiview stereo correspondence data set demonstrate that DERF outperforms state of the art methods for both hand-crafted and learned descriptors, while remaining robust and being much faster to compute.**

***Index Terms*— Computer vision, image matching, local descriptors, wavelets, ganglion cells.**

Manuscript received November 7, 2014; revised January 24, 2015; accepted February 16, 2015. Date of publication March 6, 2015; date of current version April 15, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2010CB327902, in part by the National Natural Science Foundation of China under Grant 61273263 and Grant 61202237, in part by the Beijing Municipal Natural Science Foundation under Grant 4142032, in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20121102120016, in part by the Joint Project through the LIA 2MCSI Laboratory between the Group of Ecoles Centrales and Beihang University, in part by the Fundamental Research Funds for the Central Universities, and in part by the Australian Research Council under Project DP-140102164 and Project FT-130101457. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ling Shao.

D. Weng and Y. Wang are with the State key laboratory of Virtual Reality Technology and Systems, School of Computer Science and Technology, Beihang University, Beijing 100191, China (e-mail: daweiweng0204@gmail.com; yhwang@buaa.edu.cn).

M. Gong and D. Tao are with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

H. Wei is with the Laboratory of Cognitive Model and Algorithm, School of Computer Science, Fudan University, Shanghai 200438, China (e-mail: weihui@fudan.edu.cn).

D. Huang is with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: dhuang@buaa.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2409739

I. Introduction

FEATURE descriptors play a fundamental role in many computer vision problems, including object and scene recognition [1]–[4], structure from motion [5], action recognition [6], [7], pedestrian detection [8], image retrieval [9], person reidentification [10] and face recognition [11]. The real challenge for descriptors is to describe key points with distinctive, robust, and efficient representations which are invariant to scale, rotation, and affine transformation [12]–[15].

Various hand-crafted local descriptors have been proposed, and we can roughly classify them into three categories: image gradient-based methods, such as Scale Invariant Feature Transform (SIFT) [16], Speeded Up Robust Features (SURF) [17], Gradient Location and Orientation Histogram (GLOH) [18], Principal Component Analysis SIFT (PCA-SIFT) [19], DAISY [20], Geometric Blur [21] and HSOG [22]; spatial frequency-based methods, such as Rotational Image [23], Shape Contexts [24], Directionlets [25], and Steerable Filters [26]; and differentiation and moment-based methods, such as Differentiation Invariants [27], and Moment Invariants [28].

Image gradient-based methods especially have received intense attention in the community due to their promising performance in a variety of applications. SIFT and GLOH should owe much of their strength to the use of gradient orientation histograms, which are relatively robust to distortions; however, they are both computationally demanding. Compared to SIFT, SURF [17] attains much faster speed by describing key points with the response of a few Haar-like filters, but SURF does away with SIFT's spatial weighting scheme, and all gradients over the local region contribute equally to their respective bins, which introduces damaging artifacts when used for dense matching. To alleviate this problem, DAISY [20] replaces the weighted sums of gradient norms with convolutions of the gradients in specific directions with several Gaussian filters, making DAISY much faster with no performance loss. The gradient-based methods are inspired by the biological model introduced by Edelman et. al. [29], but use different computational mechanisms to allow for small positional shift. In this paper, we proposed a method in which we model the response and distribution properties of parvocellular-projecting ganglion cells (P-GCs) in the primate retina. This method develops the gradient-based methods and theoretically explains the effectiveness of our method.

In the human visual system, invariant feature extraction is one of the most important information

processing tasks, and is also a common characteristic of senior cortex cells in the process of information integration. In the information processing mechanism of the human visual system and the corresponding physiological structure, preliminary feature representation from the retina via the Lateral Geniculate Nucleus (LGN) to cortical simple cells provides effective coding for invariant feature extraction. The ganglion cell (GC) is the last cell layer of the primate retina and produces the retinal coding. Given the neural circuit between a GC and its receptive field (RF) [30], [31], as well as the electrophysiological characteristics [32] of GCs, GCs not only function as a preparatory step for the detection of boundaries by subsequent cells, but also play a crucial role in sampling the whole visual field accurately and efficiently. A number of computational simulations have focused on GCs. These works mainly focus on edge or contour detection [33] and image enhancement [34]. Some recent studies have modeled GC and its RF for image representation [35], [36]. In [36], the whole retina is modeled to construct a binary descriptor to represent key points. However, they don't discriminate different GCs and visual pathways, and the convolution kernels are completely different from the RF of a ganglion cell.

An interesting and notable electrophysiological finding shows that the neural response of a single P-GC can be properly modeled by the difference of Gaussian (DoG) function at a single scale; When the distribution of all P-GCs is taken into consideration, we find that GCs are arranged roughly in rings with different radii, and that the scale of the response function and the interval between neighboring rings both increase exponentially [30], [37]. Another important finding about the RF of a P-GC is that it can be resized to a certain extent [38]–[40]. Electrical synapses are usually flexible enough to be formed or broken dynamically, which results in size-changeable RFs.

Proliferous learning techniques were mainly employed for higher level visual tasks, however, recently feature construction based on learning methods had emerged to remedy the plausible limitation of hand-crafted features that most hand-crafted features cannot adapt to new conditions [41], [42]. These methods can be roughly classified into two categories: learning low-level features [41], [43]–[45]; learning multiple levels of features including low-level feature (e.g., edges, object parts, and objects) to expect higher-level features to represent more abstract semantics and to provide more invariance to various distortions, i.e. deep learning neural networks [42], [46]–[49]. In human visual system, the representation of retina is not controlled by brain consciousness, its representation structure is fixed and there is no learning step to adjust the representation in various cases. However, such low-level representation of retina to the visual world can sustain subsequent various complex visual tasks on the visual pathway of human. Our descriptor derived from modeling this representation outperforms the learning-based low-level features, such as [45], [50], and [51]. Coincidentally, wavelet scattering networks [52], [53] that is the representative work of deep learning and is justified mathematically has prefixed low-level feature filters-they are simply wavelet operators, hence no learning is needed at all at the stage of low-level feature.

A. Our Contribution

By mimicking the mechanisms discovered in the aforementioned findings, this paper presents Distinctive Efficient Robust Features (DERF) which are in particular derived by comprehensively modeling the response and distribution properties of the parvocellular-projecting ganglion cells (P-GCs) in the primate retina. In addition, DERF can be well interpreted by wavelet tight frame theory. Comprehensive empirical evaluations demonstrate that DERF significantly outperforms the state of the art hand-crafted descriptors, such as SIFT, HOG, and DAISY, and even performs better than the representative learning-based descriptors, such as [45], [50], and [51]. Additionally, it is robust and computationally efficient. The contribution of this paper is twofold:

- A new distinctive, efficient, and robust descriptor inspired by the modeling of the response and distribution properties of P-GCs:

We convolve gradient maps at the locations of grid points using the DoG function, the scale of which is smaller at the grid points close to the center, and increases exponentially away from it. The grid points are arranged into concentric rings, and the radial distance between the grid points on neighboring rings also increases exponentially. Additionally, we implement DoG filtering efficiently by employing separable Gaussian filtering. Lastly, convolution with a large Gaussian kernel can be obtained from several consecutive convolutions with smaller kernels. Thus, DERF is computationally efficient. The circular symmetric grid and isotropic kernels (DoG) make DERF naturally resistant to rotational perturbations. DERF was evaluated on the image matching task on the Multi-view Stereo Correspondence Dataset [45], with and without a dimension reduction step. The results show that DERF outperforms state-of-the-art descriptors, including learning-based descriptors.

- A new theoretical support of local descriptor design:

DoG is naturally a wavelet, and the structure of grid points array in our descriptor is also closely similar to the spatial sampling of wavelets. In wavelet theory, a signal f can be characterized completely by a wavelet frame; however, a tight frame can not only represent signals completely, but can also ensure that the new representation is robust to distortions. DoG wavelet itself forms a frame, and when we modulate the parameters of DERF to make the frame tighter, the performance of DERF improves accordingly. To further validate the conclusion, we design a much tighter wavelet by adjusting the DoG wavelet that is even tighter than the Marr wavelet (second order derivate of Gaussian), called Tight Frame DoG (TF-DoG). By replacing DoG with TF-DoG in DERF, image matching performance is further improved.

The rest of the paper is organized as follows. Section 2 formalizes the design of two types of DERF based on both the response and distribution properties of P-GCs. We also briefly discuss the method of dimension reduction used in our experiments and analyze the computational complexity

of DERF. Section 3 theoretically analyzes DERF from the perspective of the wavelet frame and presents a new theoretical support of local descriptor design. We validate the conclusion by constructing a tight frame wavelet (TF-DoG) which results in better performance by DERF. In Section 4, we comprehensively evaluate DERF by conducting a large number of experiments in the image matching task on the Multi-view Stereo Correspondence Data set. Section 5 presents the performance evaluation of TF-DoG and its corresponding descriptor. Lastly, in Section 6, we summarize our research on local descriptor design.

II. DERF

Since Rodieck's (1965) DoG model of the receptive field of retinal GCs [54], an enormous amount of experimental and theoretical research has greatly advanced our understanding of the retina, and GCs in particular. Ganglion cells play a crucial role in sampling and representing the whole visual field accurately and efficiently.

Gangion cells (GCs) in the retina are classified into two categories: parvocellular-projecting GCs (P-GCs) and magnocellular-projecting GCs (M-GCs). Physiological and neuroanatomical studies of GCs in the central 10° of the primate retina (roughly 5% of the retinal area) show that the majority of these cells are P-GCs [55]–[57], and it has been proposed that P (parvocellular) and M (magnocellular) cells serve distinct functions in the visual system [58], [59], as shown in Fig. 1. The visual system in primates is divided into two pathways. The first visual pathway processes information primarily from the central region of the visual field, receives input mostly from P cells, and forms high spatial resolution information (meaningful in identifying objects) to the inferior temporal cortex. The second pathway primarily processes information about the peripheral visual field, receives input mainly from M cells, and sends information with low spatial but high temporal resolution (useful in locating objects) to the posterior parietal cortex. We can therefore model the response and distribution properties of P-GCs to construct local features for visual recognition.

A. Properties of P-GCs in the Primate Retina

The human retina has a multi-layered structure in which the receptor cells (cones and rods) form the first layer for receiving a variety of stimuli and converting them into neural impulses, and ganglion cells constitute the last layer which executes the ultimate coding for these neural impulses and transmits the code to the LGN. In neuroscience, many researchers think that the middle layers can be viewed as a black box and just to form the mapping relation between a GC and its corresponding neural impulses. It was first proposed by Rodieck [54] in 1965 that this mapping between the output of a P-GC and the stimuli on its corresponding receptor cells can be modeled by the DoG function as follows:

$$\hat{\psi}(\xi,\nu) = K_c \cdot \exp(-\frac{\xi^2+\nu^2}{2r_c^2}) - K_s \cdot \exp(-\frac{\xi^2+\nu^2}{2r_s^2}) \quad (1)$$

where K_c and K_s are equal to the peak sensitivities of the center and surround of the P-GC receptive field respectively,

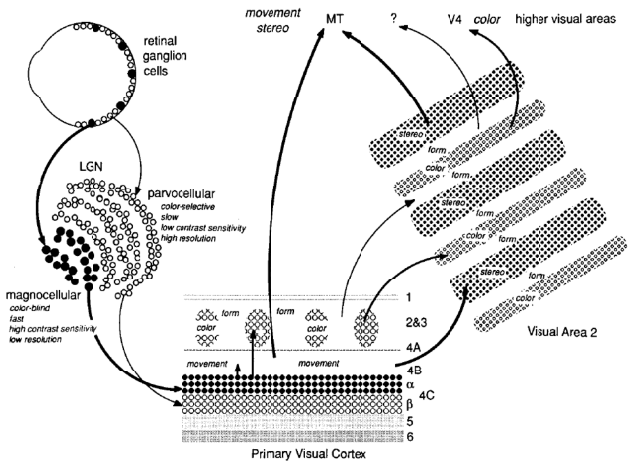


Fig. 1. Diagram of the functional segregation of the primate visual system. MT, middle temporal lobe; V4, visual area 4; LGN, lateral geniculate nucleus [59].

r_c and r_s are equal to the radii of the center and surround respectively. According to Croner's experimental results [37], the peak sensitivities of both the center and surround are inversely related to the size of each region: larger regions have less sensitivity, as described by the following equations for P-GCs across the whole retina: $K_c = 0.391 \cdot r_c^{-1.850}$ ($n = 90, r^2 = 0.67, p < 0.001$), $K_s = 0.128 \cdot r_s^{-2.147}$ ($n = 81, r^2 = 0.77, p < 0.001$). The slopes of the regression lines can be roughly viewed as constant -2. Croner also suggested that the ratio of r_s/r_c is constant across the whole retina for P-GCs. The model now becomes the following equation:

$$\hat{\psi}(\xi,\nu) = \frac{1}{2\pi r_c^2}\exp(-\frac{\xi^2+\nu^2}{2r_c^2}) - \frac{1}{2\pi r_s^2}\exp(-\frac{\xi^2+\nu^2}{2r_s^2}) \quad (2)$$

Besides the response function of a single P-GC, the following evidence from neurophysiology and anatomy has also revealed the distribution properties of P-GCs, as shown in Fig. 2.

1) Croner [37] investigated the organization of the receptive field across a wide range of retinal eccentricities and found that r_c and r_s, the radii of the center and surround regions of the receptive field of P-GC, have a steeper increase in size, especially in the range of $0° \sim 15°$ eccentricities, which is the range in which the majority of P-GCs reside.
2) The exponential increase scheme of the size of receptive field can also be reflected by the change in the size of the dendritic field of P-GCs to some extent. In the retina of both the human and macaque, it has been concluded that P-GCs, especially in the range $0° \sim 15°$, show a characteristically steeper exponential increase in dendritic field size with increasing eccentricity [30]. This can be seen as indirect evidence that r_c and r_s increase exponentially.
3) From the eccentricity $0° \sim 15°$ of the human retina, the density of P-GCs decreases exponentially [30]. This also reflects that the size of the receptive field

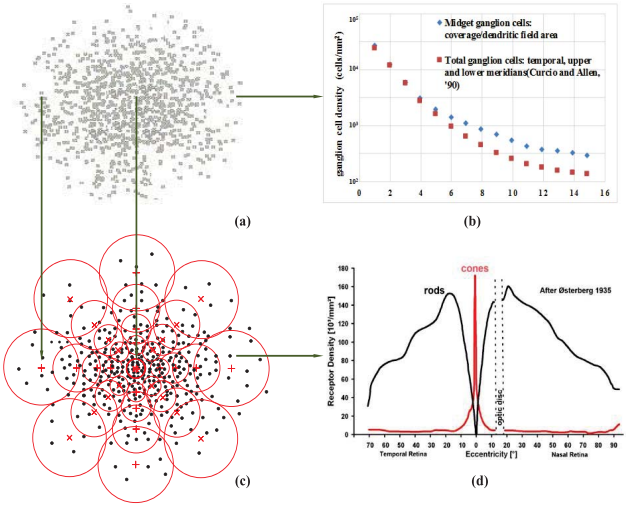


Fig. 2. The relationship between the P-GC, its receptive field, and corresponding photoreceptors cones. (a) Sketch of the distribution of P-GCs in the whole retina. (b) Calculation of ganglion cell density as a function of retinal eccentricity [30], [60]. Retinal GCs are not distributed evenly over the retina. Their density is highest close to the fovea and decreases exponentially towards the margin of the retina. (c) Sketch of the distribution of P-GC receptive fields. (d) Graph to show the densities of cones and rods along the horizontal meridian [61].

increases exponentially. As we know, if r_c increases and K_c decreases, the spatial resolution provided by the P-GC will become lower and thus the corresponding requisite sampling rate (i.e. density of cones) should also become lower. The fact that the cone density decreases exponentially at the range of eccentricity $0° \sim 15°$ effectively validates this reasoning.

In summary, we maintain that the P-GCs in the retina are distributed in roughly concentric circles, and the interval between neighboring circles increases exponentially from the midpoint to the peripheral points. In addition, the scale of the P-GC response function or the size of receptive field increases exponentially. The relationship between a P-GC, its receptive field, and the structure of the receptive fields of a P-GC array are illustrated below. Note that the exponential increase of scale is pivotal to the complete representation of the signal, which will be explained in next section.

B. Single Scale Descriptor

The proposed algorithm is summarized in the following Algorithm 1. We now detail the computation process of single scale DERF. We first compute *DoG convolved gradient orientation maps* at multiple scales and then assemble the DERF descriptor by sampling the DoG convolved gradient orientation maps. To compute the multi-scale *DoG convoluted gradient orientation maps*, we first compute the *gradient orientation maps* of H (8) directions, $G_o = (\frac{\partial I}{\partial o})^+, 1 \leq o \leq H$, where I is the input image, o is the orientation of the derivative, and $(.)^+$ is the operator such that $(a)^+ = max(a, 0)$. After the acquisition of *gradient orientation maps* $G_o, 1 \leq o \leq H$, each *gradient orientation map* is convolved $S + 1$ (6) times with Gaussian kernels of different Σ values to obtain *Gaussian convolution orientation maps* as $G_o^{\Sigma} = G_{\Sigma} * (\frac{\partial I}{\partial o})^+$

Algorithm 1 Biologically Inspired Local Descriptor DERF

Input:

Image patch around certain interest point

Output:

A feature vector to character this interest point

1: Construct grid points template: We adopt $S(5)$ concentric rings whose radii increase exponentially, and arrange evenly $T(8or12)$ sampling orientations for each ring, the scale of DoG convolution of the grid points on the ith circle equals to $\sigma_i = \eta \cdot r_i$.

2: Obtain the convolution scale of Gaussian Kernel: We use the difference of two Gaussian filtering to realize DoG filtering. $S + 1$ Gaussian scales are selected according to the scales of DoG function.

3: Compute the gradient orientation maps of $H(8)$ directions, $G_o = (\frac{\partial I}{\partial o})^+, 1 \leq o \leq H$.

4: Compute DoG convolved gradient orientation maps: for each map of H gradient orientation maps, we perform $S+1$ times Gaussian filterings, and then obtain the S scales DoG convolved maps by subtracting the latter one from the former one in the cascaded $S + 1$ Gaussian filterings. Then, according the scale of each DoG convolved gradient orientation map, we execute a Gaussian smooth of small scale.

5: Assemble the final vector: According the scale of each grid point, we extract the pixel values of each grid point in DoG convolved H gradient orientation maps to constitute the feature vector.

where G_{Σ} is a Gaussian kernel with scale Σ. Different Σs correspond to different sizes of convolution region. Lastly, for each *orientation map*, we subtract the latter from the former for each pair of neighboring *Gaussian convolution orientation maps* to obtain *DoG convolution orientation maps* as $D_o^{\Sigma_1} = G_o^{\Sigma_1} - G_o^{\Sigma_2}$, where $\Sigma_2 > \Sigma_1$.

Because we implement DoG filtering by using separable Gaussian filters, we can compute DoG convolution orientation maps at a low cost. Also, convolution with a large Gaussian kernel can be obtained from several consecutive convolutions with smaller kernels:

$$\begin{aligned} D_o^{\Sigma_1} &= G_o^{\Sigma_1} - G_o^{\Sigma_2} = G_o^{\Sigma_1} - G_{\Sigma_2} * (\frac{\partial I}{\partial o})^+ \\ &= G_o^{\Sigma_1} - G_{\Sigma} * G_{\Sigma_1} * (\frac{\partial I}{\partial o})^+ = G_o^{\Sigma_1} - G_{\Sigma} * G_o^{\Sigma_1}, \\ &\qquad \Sigma_2 > \Sigma_1, \quad \Sigma = \sqrt{\Sigma_2^2 - \Sigma_1^2}. \end{aligned}$$

Due to the isotropy of the DoG filter, the result of filtering for one interest point can be used for other interest points in an image, and the filtering for one scale can be reused for all the grid points with same scale, i.e., just one template for one scale. This computational flow, the incremental computation of the *DoG convolution orientation maps* from an input image, is summarized in Fig. 3.

After obtaining the convoluted gradient orientation maps, we construct our DERF descriptor by sampling the gradient orientation maps. Mimicking the structure of receptive fields of the P-GC array, the lattice points of our descriptor are located

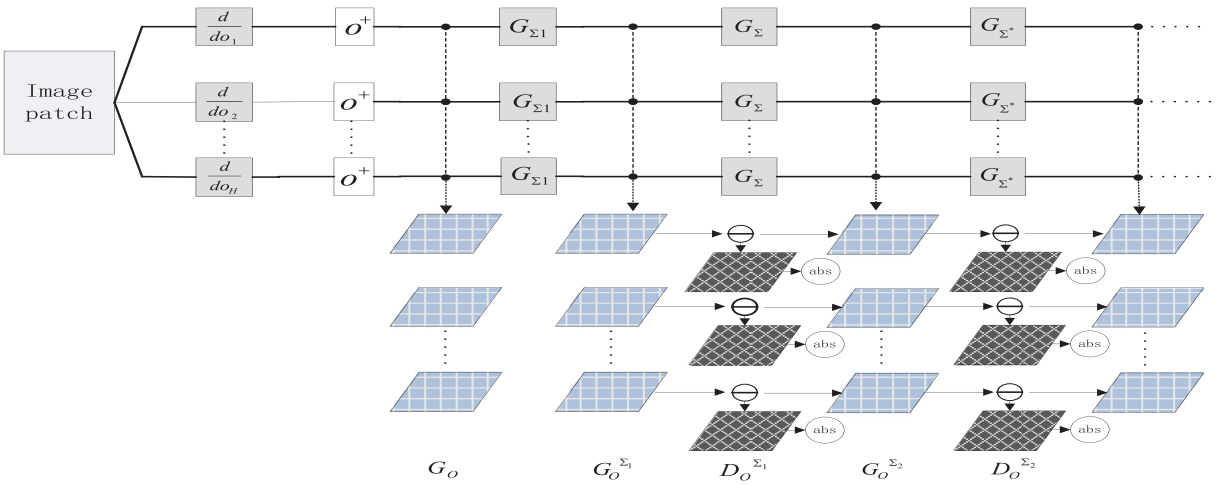
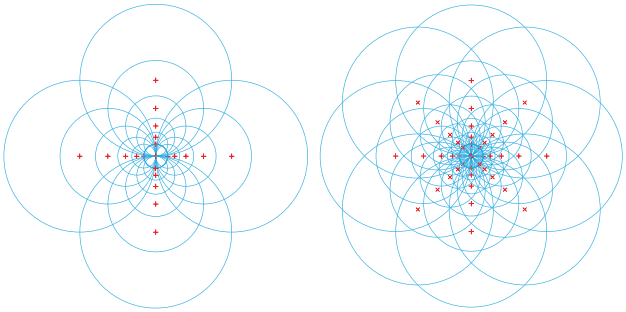


Fig. 3. Computational flow of DoG convolution orientation maps: We first compute orientation maps from the original image, which are convolved by Gaussian kernels to obtain Gaussian convolution orientation maps, and then obtain DoG convolution orientation maps by the subtraction of Gaussian convolution orientation maps. By chaining the above convolutions, the $D_o^{\Sigma i}$ can be obtained very efficiently.



(a) (b)

Fig. 4. Our descriptor: Each circle represents a region in which the radius is proportional to the standard deviation of the DoG. By overlapping the regions, we can achieve smooth transition between the regions and obtain a degree of rotational robustness. The radii of the outer regions are increased to give equal sampling along the rotational axis, which is necessary for robustness against rotation. (a) Four orientations. (b) Eight orientations.

in many concentric rings with different radii increasing in exponential manner, and the corresponding DoG convolution kernel at each grid point is smaller in the inner rings and larger in the outer rings. In our descriptor, we adopt $S(5)$ scales and $T(8\ or\ 12)$ sampling orientations for each ring. The resulting descriptor is shown in Fig. 4 and the corresponding parameters that control its shape are listed in Table I.

Formally, our descriptor consists of the values of grid points from the *DoG convolution orientation maps*. The grid points are located on concentric circles centered on an interest point, and the amount of DoG filtering is proportional to the radii of the circles, $\sigma_i = \eta \cdot r_i$. Let $h_\Sigma(\xi_0, \nu_0)$ represent the vector constructed of the values at location (ξ_0, ν_0) in the *DoG convolution orientation maps* with the same scale Σ,

$$h_\Sigma(\xi_0, \nu_0) = [D_1^\Sigma(\xi_0, \nu_0), \cdots, D_H^\Sigma(\xi_0, \nu_0)],$$

where D_1^Σ, D_2^Σ and D_H^Σ denote the DoG convolution orientation maps with the same scale and different directions. If S represents the number of layers and T the number of sampling directions on each ring, then the full descriptor $\mathcal{D}(\xi_0, \nu_0)$ for the center (ξ_0, ν_0) is defined as the concatenation of h vectors:

$$\begin{aligned}\mathcal{D}(\xi_0, \nu_0) = \Big[& h_{\Sigma_1}(\xi_0, \nu_0), \\ & h_{\Sigma_1}(l_1(\xi_0, \nu_0, R_1)), \cdots, h_{\Sigma_1}(l_T(\xi_0, \nu_0, R_1)), \\ & h_{\Sigma_2}(l_1(\xi_0, \nu_0, R_2)), \cdots, h_{\Sigma_2}(l_T(\xi_0, \nu_0, R_2)), \\ & \quad \vdots \\ & h_{\Sigma_S}(l_1(\xi_0, \nu_0, R_S)), \cdots, h_{\Sigma_S}(l_T(\xi_0, \nu_0, R_S)) \Big]^\top\end{aligned}$$

where $l_j(\xi_0, \nu_0, R)$ is the location with distance R from (ξ_0, ν_0) in the direction given by j.

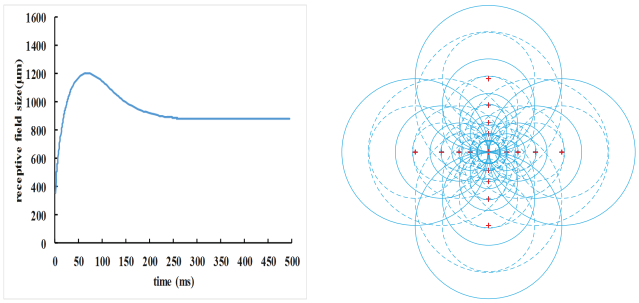
As a byproduct, using circular grid and isotropic kernels (DoG) makes our descriptor naturally resistant to rotational perturbations, and when we want to compute the descriptor of another orientation, there is no need to recompute the gradient orientation maps and the DoG convolution orientation maps; we simply rotate the sampling grid to this direction and circularly shift the vector $h_\Sigma(\xi_0, \nu_0)$ of every grid point according to the change of gradient orientation by interpolation. Lastly, a notable finding is that the average weight in the inner region, if viewing the DoG convolution as a kind of weight, is larger than that in the outer region which is accordant with the learned results in [51].

C. Multi-Scale Descriptor

Another important finding about the RF of P-GCs is that it can be resized to a certain extent [38]–[40]. Electrical synapses are usually flexible enough to be formed or broken dynamically, which results in size-changeable RFs. The degree of size modulation is as slight, as demonstrated in Fig. 5. Taking this modulation mechanism of RF into consideration, we added two neighboring scales for each grid of a single descriptor. In the single scale descriptor, the grids in

TABLE I
SHAPE PARAMETERS OF OUR DESCRIPTOR

Parameter Name	Symbol	Description
Radius	R	Distance from the center pixel to the outermost grid point.
Layer Number	S	Number of layers with different scales.
Orientation Number	T	Number of orientations at a single layer.
Grid Points Number	G	Number of all grid points used in the descriptor.
Ratio	D	The ratio between the radii of neighboring layers.



(a) (b)

Fig. 5. The dynamic size-changeable receptive field and the resulting multi-scale descriptor. (a) Receptive field size vs time plot: the receptive field first increases and then decreases to stabilize on a smaller value [38]. (b) Multi-scale descriptor: based on the single-scale descriptor (Fig. 4), two neighboring scales are added to each grid point. This structure is thus also circularly symmetrical and is beneficial for tackling rotation.

one circle share the same scale, and T (5) scales are adopted. For the grids with the smallest scale, we added the second scale to each grid, and for the grids on the outermost circle, we added the fourth scale. Except for these grids and circles, the remaining grids were added by two neighboring scales, which resulted in a total of three scales for each grid. Formally, the descriptor $\mathcal{D}(\xi_0, \nu_0)$ centered at (ξ_0, ν_0) is defined as follows:

$$\begin{aligned}
&\mathcal{D}(\xi_0, \nu_0)\\
&= \Big[h_{\Sigma_1}(\xi_0, \nu_0), h_{\Sigma_2}(\xi_0, \nu_0),\\
&\quad h_{\Sigma_1}(l_1(\xi_0, \nu_0, R_1)), h_{\Sigma_2}(l_1(\xi_0, \nu_0, R_1)), \cdots,\\
&\quad h_{\Sigma_1}(l_T(\xi_0, \nu_0, R_1)), h_{\Sigma_2}(l_T(\xi_0, \nu_0, R_1));\\
&\quad h_{\Sigma_1}(l_1(\xi_0, \nu_0, R_2)), h_{\Sigma_2}(l_1(\xi_0, \nu_0, R_2)),\\
&\quad h_{\Sigma_3}(l_1(\xi_0, \nu_0, R_2)), \ldots, h_{\Sigma_1}(l_T(\xi_0, \nu_0, R_2)),\\
&\quad h_{\Sigma_2}(l_T(\xi_0, \nu_0, R_2)), h_{\Sigma_3}(l_T(\xi_0, \nu_0, R_2));\\
&\qquad \vdots\\
&\quad h_{\Sigma_{S-1}}(l_1(\xi_0, \nu_0, R_S)), h_{\Sigma_S}(l_1(\xi_0, \nu_0, R_S)), \cdots,\\
&\quad h_{\Sigma_{S-1}}(l_T(\xi_0, \nu_0, R_S)), h_{\Sigma_S}(l_T(\xi_0, \nu_0, R_S))\Big]^{\top}
\end{aligned}$$

By employing this descriptor, a slight improvement in performance will be obtained, as demonstrated in the experimental section. In additon, when we further added the scales to each grid point, holding all the T scales, there are a little further improvements. Noticeably, the pooling regions in multi-scale case are overlapped which to some extent explains the overlap phenomenon of learned descriptor in [51]. Lastly, an interesting finding is that the structure of the multi-scale descriptor closely resembles the pattern of DoG blob detection presented by Lowe [16]. It is speculated that the human retina may also use this structure to detect blobs.

D. Discriminative Dimensionality Reduction

Many machine learning techniques have recently been applied to reduce the dimensionality of the descriptors [45], [51]. In this paper, we have employed the dimension reduction method of [51] to the DERF feature and achieved satisfactory results. In [51], Simonyan et al. proposed to reduce dimensionality as well as improve the discrimination of descriptors by learning a low-rank metric through penalising the nuclear norm of the Mahalanobis matrix. The advantage of this over other methods, such as PCA, is that the low-rank subspace is learnt discriminatively while yielding a convex problem and a globally optimal solution.

The learnt matrix W is required to project descriptors onto a lower dimensional space in which the positive and negative descriptor pairs are separated by a margin. These two requirements can be formalized by using a set of constraints:

$$W \in \mathbb{R}^{m \times n}, \quad m < n \tag{3}$$

where m is the dimensionality of the projected space and n is the descriptor dimensionality before projection:

$$d_W(\mathbf{x}, \mathbf{y}) + 1 < d_W(\mathbf{u}, \mathbf{v}), \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{P}, \quad (\mathbf{u}, \mathbf{v}) \in \mathcal{N} \tag{4}$$

where d_W is the squared L_2 distance in the projected space:

$$\begin{aligned}
d_W(\mathbf{x}, \mathbf{y}) &= \parallel W\phi(\mathbf{x}) - W\phi(\mathbf{y}) \parallel_2^2\\
&= (\phi(\mathbf{x}) - \phi(\mathbf{y}))^T W^T W(\phi(\mathbf{x}) - \phi(\mathbf{y}))\\
&= \theta(\mathbf{x}, \mathbf{y})^T A \theta(\mathbf{x}, \mathbf{y}),
\end{aligned} \tag{5}$$

where $\theta(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y})$, and $A = W^T W$ is the Mahalanobis matrix.

To guarantee the convexity of constraints (4) and (5), the optimisation is performed over the convex cone of positive semi-definite matrices $A \in \mathbb{R}^{n \times n}$, $A \succeq 0$, instead of W. The rank constraint on W (3) can be equivalently transformed into a rank constraint on A. However, the direct optimisation over $rank(A)$ is not tractable due to its nonconvexity. The convex relaxation of $rank(A)$ is the nuclear norm $\|A\|_*$ which is defined as the sum of singular values of A. Using the above soft formulations, the non-smooth *convex* objective for learning A is obtained:

$$\begin{aligned}
\arg\min_{A \succeq 0} \sum_{\substack{(\mathbf{x}, \mathbf{y}) \in \mathcal{P}\\ (\mathbf{u}, \mathbf{v}) \in \mathcal{N}}} &\mathcal{L}(\theta(\mathbf{x}, \mathbf{y})^T A \theta(\mathbf{x}, \mathbf{y})\\
&-\theta(\mathbf{u}, \mathbf{v})^T A \theta(\mathbf{u}, \mathbf{v})) + \mu_* \parallel A \parallel_*,
\end{aligned} \tag{6}$$

where $\mathcal{L}(z) = \max\{z + 1, 0\}$ is the hinge loss, and the parameter $\mu_* > 0$ trades off the empirical ranking loss versus the dimensionality of the projected space. Lastly, to optimise this objective efficiently, an effective stochastic learning technique, Regularized Dual Averaging (RDA) [62], is employed. We applied the dimension reduction method to our DERF feature and obtained better results than the original DERF and other features, which shows the effectiveness of the DERF feature.

E. Computational Complexity Analysis

In this section, we briefly review the complexity analysis of DERF. We analyze the computational complexity in the case of dense matching rather than for matching based on interest points, which provides a more comprehensive understanding of DERF. As described in Table I, the parameters of DERF include the number of gradient orientations H, the number of layers with different scales S, and the number of orientations at a single layer T. Assuming that the image has P pixels, we first need to compute the gradient orientation maps. In this stage, following the DAISY method, we obtain the gradients of other directions by rotating the horizontal and vertical gradients.

$$G_\theta = (\cos\theta \frac{\partial I}{\partial x} + \sin\theta \frac{\partial I}{\partial y})^+$$

For the computation of the horizontal and vertical gradients, we need to perform two $1D$ convolutions with kernels $[1, -1]$ and $[1, -1]^T$, which requires $2P$ additions. The other direction gradients can then be computed according to the above formula with $2P$ multiplications and P additions for each orientation. In total, $2P \times H - 4P$ multiplications and $P \times H$ additions are required for the computation of the gradient orientation maps.

To obtain DoG filterings of S scales, we perform a cascade of Gaussian filterings of $S + 1$ scales. For each scale, we need H Gaussian convolutions because of H gradient orientation maps, each is done as two successive $1D$ convolutions instead of a single $2D$ convolution, taking advantage of the separability of Gaussian kernels. Lastly, for each sale, we execute H subtractions of two smoothing gradient orientation maps, which requires $P \times H$ subtractions in total. Therefore, we require $(S + 1) \times H \times 2$ $1D$ Gaussian convolutions, and $P \times H \times S$ subtractions to obtain the DoG convolution orientation maps. These DoG convolution orientation maps are sampled at $S \times T + 1$ locations for every pixel to assemble the feature vector. Note that for the descriptors of orientations other than 0°, an additional circular shifting operation is required.

To summarize, computing all the descriptors of an image requires $2H \times (S + 1)$ $1D$ Gaussian convolutions, $(S \times T + 1) \times P$ samplings, $2P \times H - 4P$ multiplications, and $P \times H \times (S + 1)$ additions (or subtractions). In Table II, we show the computation time of all the descriptors for various sized images in a dense matching case with MATLAB code, in which the interest points are sampled every 8 pixels.

TABLE II
COMPUTATION TIME IN SECONDS ON AN ALIENWARE 14 LAPTOP

Image Size	Our Descriptor
800×600	2.5153
1024×768	4.0930
1280×960	6.6381

III. ANALYSIS FOR DERF AND PARAMETER SELECTION

The DoG function has numerous excellent characteristics, such as smoothness, circular symmetry, and compact support, and has been widely used in edge detection, image enhancement, and blob detection. Intrinsically, DoG is also a wavelet, and thus can achieve reconstruction of signal through wavelet transformation. In the modeling of descriptors, many cues, such as the exponential discretization of the scale of DoG, fit the theory of the discretized wavelet well. We verified that when the parameters of our descriptor are assigned according to the tight frame theory of discretized wavelets, the performance of our descriptor improves largely.

A. The Properties of DoG

In our descriptor, the scale (i.e. the size of the receptive field) of DoG functions is discrete and limited, therefore these elemental functions are just the discretized DoG wavelets. Usually, the discretized wavelet is obtained by discretizing the corresponding continuous wavelet as follows:

$$\psi_{m,n,k,l}(x, y) = a_0^{-m} \psi_{\theta_l}(a_0^{-m}(x - nb_0a_0^m), a_0^{-m}(y - kb_0a_0^m)) \tag{7}$$

where

$$\psi_{\theta_l}(x, y) = \psi(x\cos(l\theta_0) + y\sin(l\theta_0), -x\sin(l\theta_0) + y\cos(l\theta_0)) \tag{8}$$

is the rotated version of the mother wavelet ψ, θ_0 denotes the step size of the angular rotation, l is the index of the rotation steps, $b_0 > 0$ is the unit spatial interval, $n, k \in Z$ are the indices of shift steps along x and y directions respectively, and $a_0^m, m \in Z$ is the dilation in scale.

Due to the circular symmetry of DoG, $\psi_{\theta_l}(x, y)$ is equal to $\psi(x, y)$. Therefore, in our descriptor, we only need to use one convolution function for one scale at each grid point to largely reduce the dimensionality of DERF. As seen in (8), the scale discretization is exponential a_0^m, which is deemed to be the most feasible and efficient manner for the reconstruction of the signal by discrete wavelet transform [63]. Theoretically, the value of a_0 is not restricted, beyond $a_0 > 1$. In practice, however, it is very convenient to have $a_0 = 2$. For the spatial shift interval, $a_0^{-m}(x - nb_0a_0^m)$ implies that when scale $a = a_0^m$, the unit spatial interval is equal to $b_0a_0^m$, which ensures that the discretized wavelets at level m cover the whole 2D surface in the same way as $a = a_0$. This means that a larger scale corresponds to a larger unit spatial interval as well. In our descriptor, the grid points are located on concentric circles and the interval between neighboring rings

also increases exponentially. However, in DERF, each grid point has only one scale in the single scale case and has at most 3 scales in the multi-scale case, while the outer points only have larger scales. Therefore, the structure of our descriptor is not completely consistent with the spatial domain wavelet, which we call a wavelet-like structure.

Can signal f be reconstructed in a numerically stable way from the $<f, \psi_{m,n,k,l}>$? The numerically stable reconstruction is only possible if the $\psi_{m,n,k,l}$ constitute a frame. The concept of frame was first introduced by Duffin and Schaeffer in their research on nonharmonic Fourier series [64]. For one dimension case, it is hypothesized that $(\psi_i)_{i\in J} \in$ Hilbert space $\mathcal{H}$, if there exists $A > 0$, $B < \infty$, $\forall f \in \mathcal{H}$, the following equation always holds:

$$A \parallel f \parallel^2 \leq \sum_{i\in J} |<f, \psi_i>|^2 \leq B \parallel f \parallel^2 \tag{9}$$

then $(\psi_i)_{i\in J} \in$ forms a frame, A and B are the lower bound and upper bound of the frame, respectively. B/A denotes the tightness of the frame, and when $B = A$, the frame becomes a tight frame. If $\psi_{m,n,k,l}$ is a frame, then f can be reconstructed in a numerically stable way from $T^{wav}(f)$ with the dual frame $\tilde{\psi}$ [65]:

$$\begin{aligned} f &= \sum_{m,n,k,l} <f, \psi_{m,n,k,l}> \tilde{\psi}_{m,n,k,l} \\ &= \sum_{m,n,k,l} <f, \tilde{\psi}_{m,n,k,l}> \psi_{m,n,k,l} \end{aligned} \tag{10}$$

Unfortunately, in principle, we have to compute infinite terms for a common frame. In practice, it is therefore especially advantageous to work with frames which are almost tight, i.e. $\frac{B}{A} - 1 \ll 1$. We can can avoid complications with the dual frame; moreover, we can achieve high quality reconstruction of arbitrary f. When the frame is almost tight, the signal f can be relatively accurately reconstructed with limited terms by using the following equation:

$$f \approx \frac{2}{A+B} \sum_{m,n,k,l} <f, \psi_{m,n,k,l}> \psi_{m,n,k,l}. \tag{11}$$

When the frame is tight, in particular, the reconstruction is more accurate with following formula:

$$f = \frac{1}{A} \sum_{m,n,k,l} <f, \psi_{m,n,k,l}> \psi_{m,n,k,l} \tag{12}$$

As illustrated above, the frames are redundant. It was noticed very early by Morlet [65] that this redundancy also leads to robustness, in the sense that even if wavelet coefficients are computed with low precision, we can still reconstruct the signal with comparatively much higher precision. This redundancy can also bring robustness to our descriptor and make our descriptor character the image distinctively and robustly. In addition, it was pointed out in [66] that suboctave sampling will make the frame tighter. This is partially due to the fact that the phase space lattice, which is now a superposition of N lattices, is denser than the lattice whose translation interval is strictly proportional to the scale of the wavelet. The 2D version of the fractionally dilated wavelets is

$$\psi^{\eta}_{\theta_l}(x, y) = 2^{-\frac{2\eta}{N}} \psi_{\theta_l}(2^{-\frac{\eta}{N}}x, 2^{-\frac{\eta}{N}}y), \quad \eta = 0, \cdots, N-1. \tag{13}$$

with its fourier transform equal to

$$\hat{\psi}^{\eta}_{\theta_l}(\xi, \nu) = \hat{\psi}_{\theta_l}(2^{\eta/N}\xi, 2^{\eta/N}\nu), \quad \eta = 0, \cdots, N-1. \tag{14}$$

where N is the number of sampling frequency steps per octave, η is the index of the frequency steps per octave, and $\psi(x, y)$ is the mother wavelet function. It is known that in the visual cortex, the frequency space is also sampled every half or every third of an octave [67]. Thus, we adopted the fractionally dilated DoG wavelet as well in our descriptor.

Because the DoG wavelet is separable and can therefore save considerable computation time in two or more dimensions, it is often used to approximate the second derivate of Gaussian (LoG, also called Mexican hat wavelet) when the ratio k of standard deviations of two Gaussian kernels is roughly equal to 1.6. Coincidentally, when $k = 1.6$, our descriptor also reached better results, as demonstrated in the experimental section. This may partially be due to the fact that the Mexican hat wavelet is much tighter than the DoG wavelet [65]. The Mexican hat wavelet is an almost tight frame with the ratio $B/A = 1.083$, when $N = 1, b_0 = 0.75$. However, this wavelet is not separable, and in many applications, such as blob detection and automatic scale selection [68], it is replaced by DoG for efficiency. In the next section, we describe how, through ingeniously modulating the DoG wavelet, we obtained a tighter wavelet than LoG, which we have named tight frame DoG (TF-DoG). This achieves better matching performance, and the results verify our view that tight frame is beneficial for feature descriptors.

B. Tight Frame DoG

In this section, we will introduce a new wavelet, TF-DoG, which is obtained by ingeniously modulating the DoG wavelet. It is tighter than the Mexican hat wavelet, the second derivate of Gaussian. In the experimental section, we examined the frame bounds of TF-DoG under various sampling schemes of phase space. We applied TF-DoG to our descriptor to replace DoG and achieved better results. The success of TF-DoG validates our conclusion that making the wavelet tighter improves the performance of our descriptor. The derived wavelet is as follows,

$$\begin{aligned} &\psi(x, y, \sigma, \theta) \\ &= \frac{k\sqrt{\pi(k^2+1)}}{\pi\sigma(k^2-1)} \\ &\cdot \Big[e^{-\frac{x^2+y^2}{2\sigma^2}} \cdot (e^{i(\frac{1}{\kappa\sigma}\cdot\cos\theta\cdot x + \frac{1}{\kappa\sigma}\cdot\sin\theta\cdot y)} - e^{-\frac{1}{2\kappa^2}}) \\ &\quad - e^{-\frac{k^2(x^2+y^2)}{2\sigma^2}} \cdot (e^{i(\frac{1}{\kappa\sigma}\cdot\cos\theta\cdot x + \frac{1}{\kappa\sigma}\cdot\sin\theta\cdot y)} - e^{-\frac{1}{2k^2\kappa^2}}) \Big] \end{aligned} \tag{15}$$

where θ is the wavelet orientation in radians, $(1/\sigma)$ represents the standard deviation of Gaussian function, k represents the

ratio of two Gaussian standard deviations, κ is a fixed constant to control the spatial frequency bandwidth. The above wavelet is centered at $(x = 0, y = 0)$.

We initiate our derivation with formula (2), which is viewed as the initial frequency domain function in our work. Next, we execute normalization to make its corresponding spatial domain form satisfy $<\psi, \psi> = 1$. According to the Parseval equation $<\psi(x, y), \psi(x, y)> = \frac{1}{(2\pi)^2} \cdot <\hat{\psi}(\xi, \nu), \hat{\psi}(\xi, \nu)>$ and analyzing position (x_0, y_0), we arrive at

$$\hat{\psi}(\xi, \nu, x_0, y_0, \sigma) = \frac{4\pi\sigma k\sqrt{\pi(k^2+1)}}{k^2-1} \cdot \left[\frac{1}{2\pi\sigma^2}\left(e^{-\frac{\xi^2+\nu^2}{2\sigma^2}} - \frac{1}{k^2}e^{-\frac{\xi^2+\nu^2}{2k^2\sigma^2}}\right)\right] \cdot e^{-i(\xi x_0+\nu y_0)} \quad (16)$$

We shift its center to (ξ_0, ν_0),

$$\hat{\psi}(\xi, \nu, \xi_0, \nu_0, x_0, y_0, \sigma) = \frac{4\pi\sigma k\sqrt{\pi(k^2+1)}}{k^2-1} \cdot \frac{1}{2\pi\sigma^2}\left[e^{-\frac{(\xi-\xi_0)^2+(\nu-\nu_0)^2}{2\sigma^2}} - \frac{1}{k^2}e^{-\frac{(\xi-\xi_0)^2+(\nu-\nu_0)^2}{2k^2\sigma^2}}\right] \cdot e^{-i(\xi x_0+\nu y_0)} \quad (17)$$

In the sampling for spatial frequency domain, we hypothesize that at point (ξ_0, ν_0) of 2D frequency plane, the rotation parameter θ is assigned to $arctan\frac{\nu_0}{\xi_0}$, the analyzing point position $(x_0 = 0, y_0 = 0)$, then (17) is converted to the following formula,

$$\hat{\psi}(\xi, \nu, \xi_0, \nu_0, \sigma, \theta) = \frac{4\pi\sigma k\sqrt{\pi(k^2+1)}}{k^2-1} \cdot \left\{\frac{1}{2\pi\sigma^2}e^{-\frac{[(\xi-\xi_0)\cos\theta+(\nu-\nu_0)\sin\theta]^2+[(\nu-\nu_0)\cos\theta-(\xi-\xi_0)\sin\theta]^2}{2\sigma^2}} - \frac{1}{2\pi k^2\sigma^2}e^{-\frac{[(\xi-\xi_0)\cos\theta+(\nu-\nu_0)\sin\theta]^2+[(\nu-\nu_0)\cos\theta-(\xi-\xi_0)\sin\theta]^2}{2k^2\sigma^2}}\right\} \quad (18)$$

To make the size of the frequency window change dynamically according to the frequency center (ξ_0, ν_0), we impose $\frac{\sigma}{\omega_0} = \kappa$, where κ is set according to the requirement for spatial frequency bandwidth,

$$\hat{\psi}(\xi, \nu, \sigma, \theta) = \frac{2k\sqrt{\pi(k^2+1)}}{(k^2-1)\sigma} \cdot \left\{e^{-\frac{[(\xi-\frac{\sigma}{\kappa}\cdot\cos\theta)\cos\theta+(\nu-\frac{\sigma}{\kappa}\sin\theta)\sin\theta]^2+[(\nu-\frac{\sigma}{\kappa}\sin\theta)\cos\theta-(\xi-\frac{\sigma}{\kappa}\cos\theta)\sin\theta]^2}{2\sigma^2}} - \frac{1}{k^2}e^{-\frac{[(\xi-\frac{\sigma}{\kappa}\cos\theta)\cos\theta+(\nu-\frac{\sigma}{\kappa}\sin\theta)\sin\theta]^2+[(\nu-\frac{\sigma}{\kappa}\sin\theta)\cos\theta-(\xi-\frac{\sigma}{\kappa}\cos\theta)\sin\theta]^2}{2k^2\sigma^2}}\right\} \quad (19)$$

In terms of wavelet theory, admissible wavelets are functions having zero mean, which implies that $\|\hat{\psi}(\omega = 0)\| = 0$ is a necessary condition; otherwise the norm will become infinite in the measure of $\frac{d\omega}{\omega}$ as $\omega \to 0$. The d.c. response can be computed with $\xi = 0$ and $\nu = 0$,

$$\hat{\psi}(\xi = 0, \nu = 0; \sigma) = \frac{4\pi\sigma k\sqrt{\pi(k^2+1)}}{k^2-1}\left[\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2\kappa^2}} - \frac{1}{2\pi k^2\sigma^2}e^{-\frac{1}{2k^2\kappa^2}}\right] \quad (20)$$

Apparently, the d.c. response is not equal to zero. A family of admissible wavelets is obtained by subtracting the d.c. response,

$$\hat{\psi}(\xi, \nu, \sigma, \theta) = \frac{4\pi\sigma k\sqrt{\pi(k^2+1)}}{k^2-1} \cdot \left[\frac{1}{2\pi\sigma^2}\left(e^{-\frac{(\xi-\frac{\sigma}{\kappa}\cos\theta)^2+(\nu-\frac{\sigma}{\kappa}\sin\theta)^2}{2\sigma^2}} - e^{-\frac{\xi^2+\nu^2+(\frac{\sigma}{\kappa})^2}{2\sigma^2}}\right) - \frac{1}{2\pi k^2\sigma^2}\left(e^{-\frac{(\xi-\frac{\sigma}{\kappa}\cos\theta)^2+(\nu-\frac{\sigma}{\kappa}\sin\theta)^2}{2k^2\sigma^2}} - e^{-\frac{\xi^2+\nu^2+(\frac{\sigma}{\kappa})^2}{2k^2\sigma^2}}\right)\right] \quad (21)$$

Its inverse Fourier transform equals

$$\psi(x, y, \sigma, \theta) = \frac{\sigma k\sqrt{\pi(k^2+1)}}{\pi(k^2-1)} \cdot \left[e^{-\frac{\sigma^2(x^2+y^2)}{2}} \cdot \left(e^{i(\frac{\sigma}{\kappa}\cdot\cos\theta\cdot x+\frac{\sigma}{\kappa}\cdot\sin\theta\cdot y)} - e^{-\frac{1}{2\kappa^2}}\right) - e^{-\frac{k^2\sigma^2(x^2+y^2)}{2}} \cdot \left(e^{i(\frac{\sigma}{\kappa}\cdot\cos\theta\cdot x+\frac{\sigma}{\kappa}\cdot\sin\theta\cdot y)} - e^{-\frac{1}{2k^2\kappa^2}}\right)\right] \quad (22)$$

To more easily assay the wavelets, we change the form of the equation by using σ to replace $\frac{1}{\sigma}$. Now, we arrive at equation (15), given at the beginning of this section. We extract the mother wavelet, which generates the whole family of modeling wavelets by rotation and dilation,

$$\psi(x, y) = \frac{k\sqrt{\pi(k^2+1)}}{\pi(k^2-1)} \cdot \left[e^{-\frac{x^2+y^2}{2}}e^{i(\frac{1}{\kappa}x)} - e^{-\frac{x^2+y^2}{2}}e^{-\frac{1}{2\kappa^2}} - e^{-\frac{k^2(x^2+y^2)}{2}}e^{i(\frac{1}{\kappa}x)} + e^{-\frac{k^2(x^2+y^2)}{2}}e^{-\frac{1}{2k^2\cdot\kappa^2}}\right] \quad (23)$$

In the scheme of phase space sampling, we select three scale steps per octave. Silverman's [67] neurophysiological finding indicates that the sampling interval in spatial frequency is about three steps per octave in a monkey's visual cortex and it is also heuristic for us to set the value of N (scale steps per octave). Also, it is important to note that we adopt the frame conditions of Tai Sing Lee [69] to verify our wavelets. The verification process is demonstrated in appendix. After careful scrutiny, we found that in some cases, a discrete family of our wavelets forms an approximately tight frame. We applied TF-DoG to our single scale descriptor and attained better results. The drawback is that TF-DoG is not circularly symmetric, and at each grid point we need K convolution kernels for each scale, rather than one kernel like DoG, so the dimensionality of this descriptor is higher. Fig. 6 demonstrates the contrast between DoG and TF-DoG on one direction.

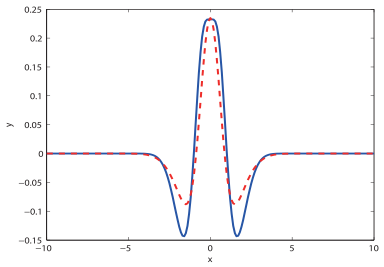
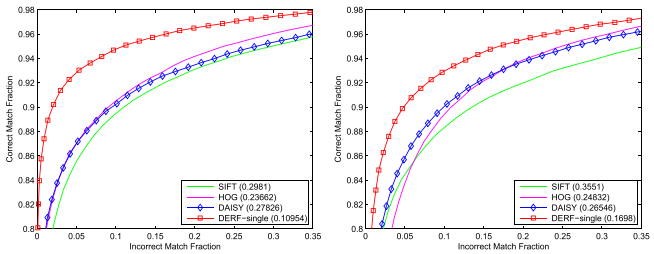


Fig. 6. Function curves of our TF-DoG (blue solid line) and DOG (red dash line) on specific direction. Since DoG is isotropic while TF-DoG has K sampling orientation for one scale, we select the equivalent orientation to compare them.



(a) (b)

Fig. 7. ROC curves for single scale DERF and other descriptors. (a) Yosemite. (b) Liberty.

IV. Performance Evaluation of DERF

In this section, we evaluate the proposed method on the dataset from actual 3D correspondences, obtained via a stereo depth map, which allows us to design descriptors to deal with non-planar transformations and illumination changes that result from viewing a true 3D scene. The dataset consists of three subsets, Yosemite, Notre Dame, and Liberty, each of which contains more than 450,000 image patches (64*64 pixels) sampled around Difference of Gaussians (DoG) feature points. 500,000 ground-truth feature pairs are generated for each subset using these patches, which contains 50 percent match pairs and 50 percent non-match pairs. Note that the scale and dominant orientation of the patches are normalized.

To compare the performance of feature descriptors, we set a threshold on the descriptor distance to generate the ROC curve and gave many results in terms of the 95 percent error rate, which is the percentage of incorrect matches on non-match pairs when 95 percent of true matches on matched pairs are obtained. Adopting the experimental scheme of [45], four training and test set combinations were used: Yosemite-Notre Dame, Yosemite-Liberty, Notre Dame-Yosemite, and Notre Dame-Liberty, in which the first of the pair is the training set. We compared our descriptor to the hand-crafted descriptors, SIFT, HOG, and DAISY, as well as the learnt descriptors, such as [45] and [51]. For hand-crafted descriptors including our DERF, we used 100,000 feature pairs of the training set to adjust their parameters and 100,000 feature pairs of the testing set to test the performance. We adjusted the scale parameter of SIFT, optimized HOG over its bin number and cell size, DAISY over the radius R, the number of layers, and the number of sampling orientations on each circle. For the learnt descriptors, all methods were trained on 500,000 feature pairs of the training set and tested on 100,000 feature pairs of the testing set.

A. Single Scale DERF

Given that the size of the patch is 64*64 in our experiment, we should take appropriate values for r_1, r_5, N. The variation of the size of the patch corresponds to the dilation of these parameters. In this case, only the values of T and S do not need to be changed. ROC curves are shown in Fig. 7, and error rates

TABLE III
Ninety-Five Percent Error Rates for Hand-Crafted Descriptors

Train	Test	DERF-single	DAISY	HOG	SIFT
Yos	ND	**9.61**	19.21	23.36	26.84
Yos	Lib	**17.21**	24.36	26.27	36.5
ND	Yos	**10.95**	27.83	23.66	29.81
ND	Lib	**16.98**	26.55	24.83	35.51
mean		**13.69**	24.49	24.53	32.17

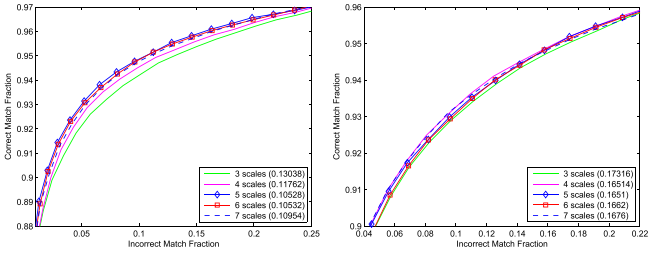
are given in Tables III and IV. We specify their dimensionality as (dim), e.g., (488) for 488-D descriptors. For our descriptor DERF in the single scale case, we obtained very good results, exceeding the performance of SIFT, HOG, DAISY, and the learnt descriptors demonstrated in [45] and [51]. Additionally, [45] pointed out that descriptors based on steerable filters (T3) usually achieve better results than those based on other gradients, at the cost of increasing computational complexity. Our result based on simple gradients is even better than the result of [45] based on T3. Recently, [51] used convex learning formulations to learn the pooling regions of the descriptor and achieved state-of-the-art performance. Although our ~~desriptor~~ does not ~~incoporate~~ the learning phase, it still achieves a slight~~er~~ better result than [51], and sets the state-of-the-art for the dataset. To guarantee a fair comparison with [45] and [51], we construct ~~three~~ types of DERF with different dimensionality, corresponding to different numbers of sampling orientations per cycle. It is also observed that initially, with the increase in grid points, the descriptor becomes better. However, when the dimensionality reaches about 800, it shows no continuous increase in matching performance.

To further explore the relationship between our descriptor and the conclusions of wavelet theory, we executed a series of other experiments. We evaluated our descriptors with different scales, i.e. different layers, and found that for any number of scale S, the best case is achieved with nearly the same r_1 and r_{last}. The best cases of different scales are shown in Fig. 8. From the results, we can see that the performance has a noticeable improvement when the scale changes from three to five, but has little improvement when the scale is larger than five. In addition, if the 1% performance improvement is disregarded, we can select four scales to further reduce the dimensionality of our descriptor.

The number of sampling orientations on each ring also has a considerable influence. As demonstrated in Fig. 9,

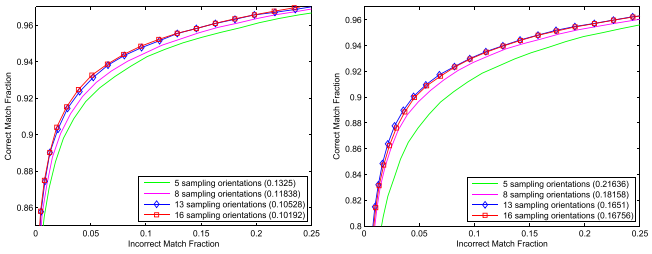
TABLE IV
NINETY-FIVE PERCENT ERROR RATES FOR LEARNT DESCRIPTORS

Train	Test	DERF-single (376)	DERF-single (536)	Simonyan et al. [51] learning ($\leq$ 640)	Simonyan et al. [51] learning ($\leq$ 384)	Brown et al. [45] learning (T3)	rootSIFT [70]
Yos	ND	9.61 (376)	**9.4 (536)**	9.49 (544)	9.88 (352)	14.43 (400)	22.06 (128)
Yos	Lib	17.21 (376)	**16.68 (536)**	17.23 (544)	17.86 (352)	20.48 (400)	29.65 (128)
ND	Yos	10.95 (376)	**10.53 (536)**	11.11 (576)	10.91 (352)	15.91 (544)	26.71 (128)
ND	Lib	16.98 (376)	**16.51 (536)**	16.56 (576)	17.02 (352)	21.85 (400)	29.65 (128)
mean		13.69	**13.28**	13.60	13.92	18.17	27.02



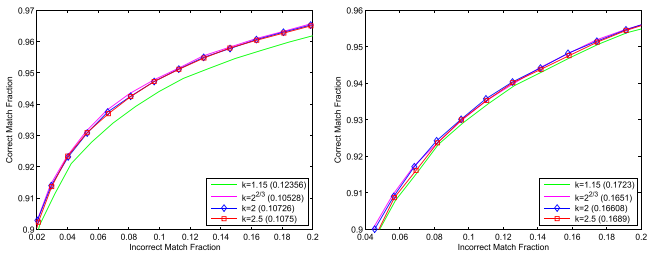
(a) (b)

Fig. 8. ROC curves for our descriptors with different scales. (a) Yosemite. (b) Liberty.



(a) (b)

Fig. 9. ROC curves for our descriptors with different numbers of sampling orientations. (a) Yosemite. (b) Liberty.



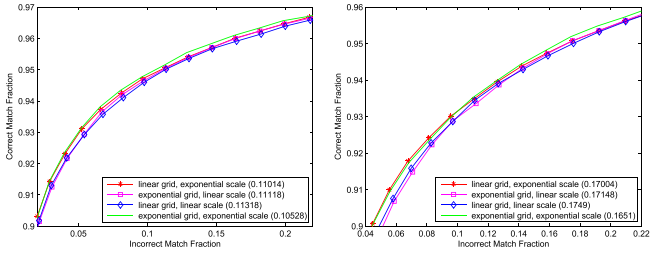
(a) (b)

Fig. 10. ROC curves for our descriptors with different values of deviations ratio k. (a) Yosemite. (b) Liberty.

the performance will show little improvement when the sampling orientation is up to 13 orientations.

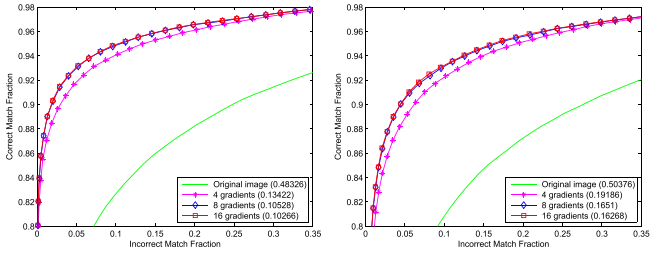
For fixed N and η, the value of the deviations ratio k of two Gaussian kernels also influences the performance slightly. A relatively optimum value exists. As described in Fig. 10, the optimum value is $k = 2^{2/3}$ where the frame of the DoG wavelet is tightest, and DoG can be viewed as the approximate surrogate of the second order derivate of Gausssian.

To show the merits of our descriptor structure, we also compare our descriptor to three other descriptor structures.



(a) (b)

Fig. 11. ROC curves for our descriptors with different grid structures and scale sampling manners. (a) Yosemite. (b) Liberty.

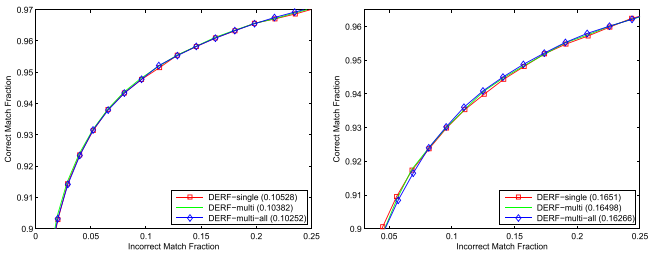


(a) (b)

Fig. 12. ROC curves for our descriptors extracting information from 4, 8, 16 orientations gradient maps, and original image. (a) Yosemite. (b) Liberty.

The first structure has the same layer number and the same r_1, r_5, and $\sigma_i = \eta \cdot r_i$ as our descriptor, but its grid is arranged with uniformly spaced circles, i.e. linear grid, linear scale. The second structure has the same layer number and the same r_1, r_5 but the grid is arranged with uniformly spaced circles, and the scale increases exponentially, i.e. linear grid, exponential scale. The third structure has the same layer number, the same r_i, and the same sampling bandwidth but the scale increases linearly, i.e. exponential grid, linear scale. Our descriptor outperforms all of them, as shown in Fig. 11 which implies that exponentially increasing grid structure and scale are not only helpful to the tightness of the frame but also to the performance of our descriptor.

To show the influence of orientation sampling, our descriptor was imposed on four orientation gradient maps, eight orientation gradient maps, 16 orientation gradient maps, and the original image, respectively, as shown in Fig. 12. From these results, we can see that although the performance improves with the increase in orientation, eight orientation gradient maps are sufficient to obtain good performance.



(a) (b)

Fig. 13. ROC curves for single scale and multi scale DERF. (a) Yosemite. (b) Liberty.

TABLE V
NINETY-FIVE PERCENT ERROR RATES FOR MULTI SCALE DERF

Train	Test	DERF-single (536)	DERF-multi (1,376)	DERF-multi-all (2,680)
Yos	ND	9.4	9.38	**9.31**
Yos	Lib	16.68	16.54	**16.38**
ND	Yos	10.53	10.38	**10.25**
ND	Lib	16.51	16.5	**16.27**
mean		13.28	13.2	**13.05**

Lastly, there are a number of other important conclusions related to the design of our descriptor. First, the scales must keep monotonically non-decreasing away from the center. Second, the grid and scales must be arranged strictly regularly, such as in a linear or exponential manner. Lastly, the sampling of intensity of grid points and scales should avoid oversampling. In the image matching tasks, we finally took the normalization method of SIFT to overcome the change of illumination, and achieved a better result. For the stereo application, we can use a normalization method like DAISY. For other applications, the normalization should probably be changed depending on the specifics of the application.

B. Multi Scale DERF

In this section, we evaluate the performance of DERF in the multi-scale case and compare it with the single scale case. ROC curves are shown in Fig. 13, and the comparison with single case is given in Table V. As discussed above, when the grid points and scales are upped to a certain extent in the single scale case, the performance of the descriptor becomes steady. However, through adding the neighboring scales to each grid point, we still attained slightly better results than the single case. To facilitate a fair comparison, the multi-scale case is constructed based on the single scale case of Column 3. We also assess the multi-all DERF where each grid point holds five scales, and find that its performance still slightly improves, while the dimensionality is much higher. Especially after dimension reduction, the performance of multi-scale DERF improves more than the single scale DERF. The success of multi-scale DERF to some extent justifies our wavelet explanation of DERF, as the grid structure and scale distribution both resemble the spacial sampling of wavelet.

C. Dimension Reduced DERF

For the dimensionality reduction experiments, we projected the single scale DERF with dimensionality 536 (the third column in Table V) and the multi-all DERF with dimensionality 2,680 (the fifth column in Table V) onto the learnt linear subspace using the method in [51]. In Table VI, we compare our results with the results of [45], [50], and [51]. Of these three methods, the best results are achieved by [51]. To facilitate a fair comparison, we learnt three kinds of dimensionality: $\leq 80 - D$, $\leq 64 - D$, $\leq 32 - D$. As can be seen, even with low-dimensional 32-D descriptors, our method outperforms methods [45] and [50]. Combining the three dimensionality cases, our descriptor is slightly better than [51]. A noteworthy observation is that the multi-scale DERF improves more than single scale DERF after dimension reduction.

V. PERFORMANCE EXAMINATION OF TF-DoG

To determine whether the modeling wavelets form a frame and how tight that frame is, we computed the frame bounds using the frame bound equations (24), (25) and the mother wavelet equation (29). Note that, if the upper bound of the unit spatial interval b_0 is 1 for the one-octave sampling scheme, which can be simply deduced from the Nyquist sampling theorem, then for fractionally dilated wavelets, the upper bound of b_0 could be pushed above 1 because the sampling lattice becomes denser. The frame bounds computed for different sampling densities are shown in Table VII.

Several general observations can be made from the above results. First, the frame becomes tighter with the increase in the number of orientations of each scale, and in the number of frequency steps per octave, but 8 orientations and $N = 3$ are sufficient to form an approximately tight frame which is tighter than the second order derivate of Gaussian. Second, the cues of Table VII show that a small number of scales form an almost tight frame, which accords with the performance of our descriptor as shown in Fig.8. Finally, for fixed value N, the values of κ and deviations ratio k also influence the tightness of the frame, which is accordant with our descriptor as well. The optimum value of k is roughly $2^{\frac{2}{3}}$, which is coincidentally accordant with the DoG wavelet. The curve of TF-DOG with $\kappa = 1.5$, $k = 2^{\frac{2}{3}}$ is demonstrated in Fig. 6.

A. Performance Examination of the Descriptor Based on TF-DoG

We assess the performance of the TF-DoG based descriptor. The TF-DoG with eight orientations per scale and $\kappa = 1.5$ $k = 2^{\frac{2}{3}}$ is used to replace the DoG wavelet of single scale DERF corresponding to the third column of Table IV. The dimensionality of this new descriptor is 3,008. As demonstrated in Table VIII, the new descriptor outperforms the DoG based descriptor which to some extent justifies our conclusion that the tighter the frame, the better the descriptor. Compared to DERF based on DoG, the drawback of this new descriptor lies in the fact that the computational complexity is higher and TF-DoG is not circularly symmetrical. However, one can construct more efficient tight frames to develop more effective descriptors in the future.

TABLE VI
NINETY-FIVE PERCENT ERROR RATES FOR DIMENSION REDUCED DESCRIPTORS

Train	Test	DERF-multi-all (2,680, ≤ 80)	DERF-single (536, ≤ 80)	DERF-single (536, ≤ 64)	DERF-single (536, ≤ 32)	Simonyan [51] (≤ 80)	Simonyan [51] (≤ 64)	Simonyan [51] (≤ 32)	Brown [45]	Trzcinski [50]
Yos	ND	**6.02 (76)**	6.32 (79)	6.6 (57)	9.35 (31)	6.82 (76)	7.11 (58)	9.99 (32)	11.98 (29)	13.73 (64)
Yos	Lib	**12.88 (76)**	13.35 (79)	13.59 (57)	15.39 (31)	14.58 (76)	14.82 (58)	16.7 (32)	18.27 (29)	21.03 (64)
ND	Yos	**8.77 (74)**	9.23 (80)	9.65 (61)	12.46 (32)	10.08 (73)	10.54 (63)	13.4 (32)	13.55 (36)	15.86 (64)
ND	Lib	**11.1 (74)**	11.59 (80)	12.02 (61)	13.42 (32)	12.42 (73)	12.88 (63)	14.26 (32)	16.85 (36)	18.05 (64)
mean		**9.69**	10.12	10.47	12.66	10.98	11.34	13.59	15.16	17.17

TABLE VII
FRAME BOUNDS OF TF-DoG FOR DIFFERENT SAMPLING DENSITIES AND PARAMETERS. K THE NUMBER OF SAMPLING ORIENTATIONS, b_0 THE UNIT SPACIAL SAMPLING INTERVAL, N THE NUMBER OF FREQUENCY STEPS PER OCTAVE, k THE DEVIATIONS RATIO, M THE RANGES OF SCALES

b_0	k	M	κ	N	K	A	B	B/A
0.75	$2^{2/3}$	-	1.5	3	4	44.9807	45.5771	1.0133
					6	63.6724	63.7468	1.0012
					8	84.7598	84.7706	1.00013
					12	127.1369	127.1466	1.00008
0.75	$2^{2/3}$	-	1.5	1	8	27.7871	28.7388	1.0342
				2		56.5035	56.5168	1.0002
				3		84.7598	84.7706	1.00013
				4		113.0141	113.0264	1.00011
0.75	$2^{2/3}$	$[-2,2]$	1.5	3	8	70.3135	80.9192	1.1508
		$[-3,3]$				80.9103	83.7918	1.0356
		$[-4,4]$				83.7822	84.5248	1.0089
		$[-5,5]$				84.5147	84.7091	1.0023
0.75	$2^{2/3}$	-	0.3	3	8	25.9442	38.0075	1.4650
			0.6			121.6443	123.4678	1.0150
			0.9			140.9314	141.1253	1.0014
			1.2			112.4692	112.5085	1.0003
			1.5			84.7598	84.7706	1.00013
0.4	$2^{2/3}$	-	1.5	3	8	84.7631	84.7673	1.00004
0.6						84.7631	84.7673	1.00004
0.75						84.7598	84.7706	1.00013
1.0						84.4911	85.0393	1.0065
1.2						82.8551	86.6753	1.0461
0.75	1.15	-	1.5	3	8	129.8081	129.8589	1.00039
	$2^{2/3}$					84.7598	84.7706	1.00013
	2					66.874	66.9892	1.0017
	2.5					56.0371	56.434	1.0071
	4					44.0162	45.7484	1.0394

TABLE VIII
NINETY-FIVE PERCENT ERROR RATES FOR LEARNT DESCRIPTORS

Train	Test	DERF-single (376)	DERF-single (536)	DERF-multi-all (2,680)	DERF-multi-all (2,680, ≤ 80)	TF-DoG DERF (single, 3,008)	TF-DoG DERF (multi-all, 15,040)	TF-DoG DERF (single, ≤ 80)
Yos	ND	9.61 (376)	9.4 (536)	9.31 (2,680)	6.02 (76)	8.89 (3,008)	8.35 (15,040)	**5.88 (78)**
Yos	Lib	17.21 (376)	16.68 (536)	16.38 (2,680)	12.88 (76)	15.67 (3,008)	14.79 (15,040)	**12.17 (78)**
ND	Yos	10.95 (376)	10.53 (536)	10.25 (2,680)	8.77 (74)	9.45 (3,008)	8.96 (15,040)	**7.95 (75)**
ND	Lib	16.98 (376)	16.51 (536)	16.27 (2,680)	11.1 (74)	15.55 (3,008)	14.66 (15,040)	**10.78 (75)**
mean		13.69	13.28	13.05	9.69	12.39	11.69	**9.2**

VI. CONCLUSION

In this paper, a new distinctive, efficient, and robust local image descriptor, DERF, is presented by modeling the response and distribution properties of the P ganglion cells (P-GCs) in the primate retina. DERF features exponential scale distribution, exponential grid structure, and circularly symmetric convolution function Difference of Gaussian (DoG), all of which are consistent with the characteristics of the ganglion cell array found in neurophysiology, anatomy, and biophysics. We implement DoG convolution efficiently by employing separable Gaussian filtering. Convolution with a large Gaussian kernel can be obtained from several consecutive convolutions with smaller kernels. Thus, DERF is computationally efficient. In addition, a new explanation for local descriptor design is presented from the perspective of wavelet tight frames: when we modulate the

parameters of our descriptor to make the frame constructed by the convolution function tighter, the performance of the descriptor improves accordingly. This is justified by designing a tight frame DoG (TF-DoG) which leads to much better performance. Compared to DERF based on DoG, the drawback of TF-DoG based descriptor lies in the fact that the computational complexity is higher and TF-DoG is not circularly symmetrical. However, one can construct more efficient tight frames to develop more effective descriptors in the future.

Appendix
Frame Bounds of TF-DoG

The frame bounds are described below:

$$A = \frac{1}{16b_0^2}\left\{ inf_{\xi,\nu\in S}\sum_{\eta=0}^{N-1}\sum_{m\in Z}\sum_{l\in Q}\frac{1}{2}\left[\left|\hat{\psi}_{\theta_l}^{\eta}(a_0^m\xi, a_0^m\nu)\right|^2 + \left|\hat{\psi}_{\theta_l}^{\eta}(-a_0^m\xi, -a_0^m\nu)\right|^2\right] - \tilde{R}\right\} \tag{24}$$

$$B = \frac{1}{16b_0^2}\left\{ sup_{\xi,\nu\in S}\sum_{\eta=0}^{N-1}\sum_{m\in Z}\sum_{l\in Q}\frac{1}{2}\left[\left|\hat{\psi}_{\theta_l}^{\eta}(a_0^m\xi, a_0^m\nu)\right|^2 + \left|\hat{\psi}_{\theta_l}^{\eta}(-a_0^m\xi, -a_0^m\nu)\right|^2\right] + \tilde{R}\right\} \tag{25}$$

where

$$\tilde{R} = \sum_{sign=+,-}\sum_{\eta=0}^{N-1}\sum_{(p,q)\in Z^2\setminus\{(0,0)\}} \times [\beta_{sign}^{\eta}(\frac{2\pi p}{b_0}, \frac{2\pi q}{b_0})\beta_{sign}^{\eta}(\frac{-2\pi p}{b_0}, \frac{-2\pi q}{b_0})]^{1/2} \tag{26}$$

and

$$\begin{aligned}\beta_{sign}^{\eta}(s,t) &= \frac{1}{4}sup_{\xi,\nu\in S}\sum_{m\in z}\sum_{l=1}^{K}|\hat{\psi}_{\theta_l}^{\eta}(a_0^m\xi, a_0^m\nu) \\ &+ sign\cdot\hat{\psi}_{\theta_l}^{\eta}(-a_0^m\xi, -a_0^m\nu)|\cdot|\hat{\psi}_{\theta_l}^{\eta}(a_0^m\xi + s, a_0^m\nu + t) \\ &+ sign\cdot\hat{\psi}_{\theta_l}^{\eta}(-a_0^m\xi - s, -a_0^m\nu - t)|\end{aligned} \tag{27}$$

$$S = \begin{cases} 0 \le \tan(\frac{\nu}{\xi}) \le \frac{2\pi}{K} \\ 1 \le \sqrt{\xi^2+\nu^2} \le a_0 \end{cases} \tag{28}$$

$\theta_0 = 2\pi/K$, K is the number of sampling orientations, $Q = 0, 1, 2, \ldots, K-1$, N is the number of frequency steps per octave. We compute our own $\hat{\psi}_{\theta_l}^{\eta}(\xi,\nu)$ occurring in the above conditions. Because we use the following discretization,

$$\hat{\psi}_{\theta_l}^{\eta}(\xi,\nu) = \hat{\psi}_{\theta_l}(2^{\eta/N}\xi, 2^{\eta/N}\nu), \quad \eta = 0, \cdots, N-1.$$

combined with equation (23), we derive that

$$\begin{aligned}\hat{\psi}_{\theta_l}^{\eta}(\xi,\nu) &= \frac{k\cdot\sqrt{\pi(k^2+1)}}{\pi(k^2-1)}\cdot\Big[2\pi e^{\gamma_1(\xi,\nu,\eta,\theta)} - 2\pi e^{\gamma_2(\xi,\nu,\eta,\theta)} \\ &\quad - \frac{2\pi}{k^2}e^{\gamma_3(\xi,\nu,\eta,\theta)} + \frac{2\pi}{k^2}e^{\gamma_4(\xi,\nu,\eta,\theta)}\Big]\end{aligned} \tag{29}$$

where

$$\begin{aligned}\gamma_1(\xi,\nu,\eta,\theta) &= -\Big((2^{\eta/N}\xi cos l\theta_0 + 2^{\eta/N}\nu\sin l\theta_0 - \frac{1}{\kappa})^2 \\ &\quad + (2^{\eta/N}\nu\cos l\theta_0 - 2^{\eta/N}\xi\sin l\theta_0)^2\Big)/2 \\ \gamma_2(\xi,\nu,\eta,\theta) &= -\Big((2^{\eta/N}\xi\cos l\theta_0 + 2^{\eta/N}\nu\sin l\theta_0)^2 \\ &\quad + (2^{\eta/N}\nu\cos l\theta_0 - 2^{\eta/N}\xi\sin l\theta_0)^2\Big)/2 \\ &\quad - \frac{1}{2\kappa^2} \\ \gamma_3(\xi,\nu,\eta,\theta) &= -\Big((2^{\eta/N}\xi\cos l\theta_0 + 2^{\eta/N}\nu\sin l\theta_0 - \frac{1}{\kappa})^2 \\ &\quad + (2^{\eta/N}\nu\cos l\theta_0 - 2^{\eta/N}\xi\sin l\theta_0)^2\Big)/(2k^2) \\ \gamma_4(\xi,\nu,\eta,\theta) &= -\Big((2^{\eta/N}\xi\cos l\theta_0 + 2^{\eta/N}\nu\sin l\theta_0)^2 \\ &\quad + (2^{\eta/N}\nu\cos l\theta_0 - 2^{\eta/N}\xi\sin l\theta_0)^2\Big)/(2k^2) \\ &\quad - \frac{1}{2k^2\kappa^2}.\end{aligned}$$

References

[1] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[2] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug. 2014.

[3] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. Cybern.*, to be published.

[4] X. Li, L. Mou, and X. Lu, "Scene parsing from an MAP perspective," *IEEE Trans. Cybern.*, to be published.

[5] M. Pollefeys *et al.*, "Visual modeling with a hand-held camera," *Int. J. Comput. Vis.*, vol. 59, no. 3, pp. 207–232, Sep. 2004.

[6] L. Liu, L. Shao, F. Zheng, and X. Li, "Realistic action recognition via sparsely-constructed Gaussian processes," *Pattern Recognit.*, vol. 47, no. 12, pp. 3819–3827, Dec. 2014.

[7] D. Tao, L. Jin, Y. Wang, and X. Li, "Rank preserving discriminant analysis for human behavior recognition on wireless sensor networks," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 813–823, Feb. 2014.

[8] M. Wang, W. Li, and X. Wang, "Transferring a generic pedestrian detector towards specific scenes," in *Proc. IEEE CVPR*, Jun. 2012, pp. 3274–3281.

[9] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[10] D. Tao, L. Jin, Y. Wang, and X. Li, "Person reidentification by minimum classification error-based KISS metric learning," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 242–252, Feb. 2015.

[11] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 980–993, Mar. 2015.

[12] R. Maani, S. Kalra, and Y.-H. Yang, "Robust volumetric texture classification of magnetic resonance images of the brain using local frequency descriptor," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4625–4636, Oct. 2014.

[13] M. Amiri and H. R. Rabiee, "RASIM: A novel rotation and scale invariant matching of local image interest points," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3580–3591, Dec. 2011.

[14] C. Ding, J. Choi, D. Tao, and L. S. Davis. (2014). "Multi-directional multi-level dual-cross patterns for robust face recognition." [Online]. Available: http://arxiv.org/abs/1401.5311

[15] X. Lu and X. Li, "Multiresolution imaging," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 149–160, Jan. 2014.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, 2006, pp. 404–417.

[18] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[19] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE CVPR*, Jun./Jul. 2004, pp. II-506–II-513.

[20] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[21] A. C. Berg and J. Malik, "Geometric blur for template matching," in *Proc. IEEE CVPR*, Dec. 2001, pp. I-607–I-614.

[22] D. Huang, C. Zhu, Y. Wang, and L. Chen, "HSOG: A novel local image descriptor based on histograms of the second-order gradients," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4680–4695, Nov. 2014.

[23] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 433–449, May 1999.

[24] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[25] D. Jayachandra and A. Makur, "Directionlets using in-phase lifting for image representation," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 240–249, Jan. 2014.

[26] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, Sep. 1991.

[27] J. J. Koenderink and A. J. van Doorn, "Representation of local geometry in the visual system," *Biol. Cybern.*, vol. 55, no. 6, pp. 367–375, Mar. 1987.

[28] L. Van Gool, T. Moons, and D. Ungureanu, "Affine/photometric invariants for planar intensity patterns," in *Proc. ECCV*, 1996, pp. 642–651.

[29] S. Edelman, N. Intrator, and T. Poggio, "Complex cells and object recognition," 1997.

[30] D. M. Dacey, "The mosaic of midget ganglion cells in the human retina," *J. Neurosci.*, vol. 13, no. 12, pp. 5334–5355, 1993.

[31] D. M. Dacey and M. R. Petersen, "Dendritic field size and morphology of midget and parasol ganglion cells of the human retina," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 89, no. 20, pp. 9666–9670, Oct. 1992.

[32] L. Chao-Yi, Z. Yi-Xiong, P. Xing, Q. Fang-Tu, T. Cheng-Quan, and X. Xing-Zhen, "Extensive disinhibitory region beyond the classical receptive field of cat retinal ganglion cells," *Vis. Res.*, vol. 32, no. 2, pp. 219–228, Feb. 1992.

[33] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[34] K. Ghosh, S. Sarkar, and K. Bhaumik, "Image enhancement by high-order Gaussian derivative filters simulating non-classical receptive fields in the human visual system," in *Pattern Recognition and Machine Intelligence*, S. Pal, S. Bandyopadhyay, and S. Biswas, Eds. Berlin, Germany: Springer-Verlag, 2005, pp. 453–458.

[35] O. Linde and T. Lindeberg, "Object recognition using composed receptive field histograms of higher dimensionality," in *Proc. ICPR*, vol. 2. Aug. 2004, pp. 1–6.

[36] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE CVPR*, Jun. 2012, pp. 510–517.

[37] L. J. Croner and E. Kaplan, "Receptive fields of P and M ganglion cells across the primate retina," *Vis. Res.*, vol. 35, no. 1, pp. 7–24, Jan. 1995.

[38] M. Kamermans, J. Hark, J. B. A. Habraken, and H. Spekreijse, "The size of the horizontal cell receptive fields adapts to the stimulus in the light adapted goldfish retina," *Vis. Res.*, vol. 36, no. 24, pp. 4105–4119, Dec. 1996.

[39] R. Shapley, "Retinal physiology: Adapting to the changing scene," *Current Biol.*, vol. 7, no. 7, pp. R421–R423, Jul. 1997.

[40] Y. Li, H. Li, H. Q. Gong, P. Liang, and P. Zhang, "Characteristics of receptive field encoded by synchronized firing pattern of ganglion cell group," *Acta Biophys. Sin.*, vol. 27, no. 3, pp. 211–221, 2011.

[41] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multiobjective genetic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[42] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[43] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1465–1479, Sep. 2006.

[44] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.

[45] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, Jan. 2011.

[46] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. (2014). "PCANet: A simple deep learning baseline for image classification?" [Online]. Available: http://arxiv.org/abs/1404.3606

[47] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE CVPR*, Jun. 2012, pp. 3258–3265.

[48] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2303–2308, Dec. 2014.

[49] Y. Yuan, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[50] T. Trzcinski, M. Christoudias, V. Lepetit, and P. Fua, "Learning image descriptors with the boosting-trick," in *Proc. NIPS*, 2012, pp. 278–286.

[51] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1573–1585, Aug. 2014.

[52] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[53] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," in *Proc. IEEE CVPR*, Jun. 2013, pp. 1233–1240.

[54] R. W. Rodieck, "Quantitative analysis of cat retinal ganglion cell response to visual stimuli," *Vis. Res.*, vol. 5, no. 12, pp. 583–601, Dec. 1965.

[55] G. Peter, "Identification of cone mechanisms in monkey ganglion cells," *J. Physiol.*, vol. 199, no. 3, pp. 533–547, Dec. 1968.

[56] F. M. D. Monasterio and P. Gouras, "Functional properties of ganglion cells of the rhesus monkey retina," *J. Physiol.*, vol. 251, no. 1, pp. 167–195, Sep. 1975.

[57] E. Kaplan and R. M. Shapley, "X and Y cells in the lateral geniculate nucleus of macaque monkeys," *J. Physiol.*, vol. 330, no. 1, pp. 125–143, Sep. 1982.

[58] W. H. Merigan, L. M. Katz, and J. H. Maunsell, "The effects of parvocellular lateral geniculate lesions on the acuity and contrast sensitivity of macaque monkeys," *J. Neurosci.*, vol. 11, no. 4, pp. 994–1001, Apr. 1991.

[59] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: Anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, May 1988.

[60] C. A. Curcio and K. A. Allen, "Topography of ganglion cells in human retina," *J. Comparative Neurol.*, vol. 300, no. 1, pp. 5–25, Oct. 1990.

[61] G. Østerberg, *Topography of the Layer of Rods and Cones in the Human Retina*. Nyt Nordisk Forlag, 1935.

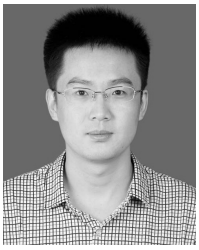
[62] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2116–2124.

[63] I. Daubechies *et al.*, *Ten Lectures on Wavelets*, vol. 61. Philadelphia, PA, USA: SIAM, 1992.

[64] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, no. 2, pp. 341–366, Mar. 1952.

[65] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961–1005, Sep. 1990.

[66] A. Grossmann, R. Kronland-Martinet, and J. Morlet, "Reading and understanding continuous wavelet transforms," in *Wavelets*, J.-M. Combes, A. Grossmann, and P. Tchamitchian, Eds. Berlin, Germany: Springer-Verlag, 1989, pp. 2–20.
[67] M. S. Silverman, D. H. Grosof, R. L. D. Valois, and S. D. Elfar, "Spatial-frequency organization in primate striate cortex," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 86, no. 2, pp. 711–715, Jan. 1989.
[68] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE ICCV*, vol. 2. Sep. 1999, pp. 1150–1157.
[69] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, Oct. 1996.
[70] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2911–2918.



Dawei Weng received the B.S. degree in computer science from the Shandong University of Finance and Economics, Jinan, China, in 2006, and the M.S. degree in computer science from Shandong University, Jinan, in 2009. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, Beijing, China. His current research interests include biometrics, neuro-inspired computer vision, and machine learning.

Yunhong Wang (M'98) received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively, all in electronics engineering. She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is also the Director of Laboratory of Intelligent Recognition and Image Processing with the Beijing Key Laboratory of Digital Media. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing. Her research results have published at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-IP, T-IFS, CVPR, ICCV, ECCV, etc.

Mingming Gong received the B.S. degree in electrical engineering from Nanjing University, Nanjing, China, in 2009, and the M.S. degree in communications and information system from the Huazhong Unviersity of Science and Technology, Wuhan, China, in 2012. He is currently pursuing the Ph.D. degree with the School of Software, University of Technology, Sydney, Australia. His research interests include causal inference, transfer learning, deep neural networks, and invariant feature learning.

Dacheng Tao (F'15) is currently a Professor of Computer Science with the Centre for Quantum Computation and Intelligent Systems, and the Faculty of Engineering and Information Technology, University of Technology, Sydney. He mainly applies statistics and mathematics to data analytics and his research interests spread across computer vision, data science, image processing, machine learning, neural networks and video surveillance. His research results have expounded in one monograph and 100+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award in IEEE ICDM'07, the Best Student Paper Award in IEEE ICDM'13, and the 2014 ICDM 10 Year Highest-Impact Paper Award.

Hui Wei received the Ph.D. degree from the Department of Computer Science, Beijing University of Aeronautics and Astronautics, in 1998. From 1998 to 2000, he was a Post-Doctoral Fellow with the Department of Computer Science and the Institute of Artificial Intelligence, Zhejiang University. Since 2000, he has been with the Department of Computer Science and Engineering, Fudan University. His research interests include neuro-inspired artificial intelligence and cognitive science.

Di Huang (S'10–M'11) received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from the Ecole Centrale de Lyon, Lyon, France, in 2011. He joined the Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, as a Faculty Member. His current research interests include biometrics, in particular, on 2D/3D face analysis, image/video processing, and pattern recognition.