Neurocomputing 119 (2013) 425–433

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

A two-layer framework for appearance based recognition using spatial and discriminant influences

Qi Li [a,*], Chang-Tien Lu [b]

[a] Department of Computer Science, Western Kentucky University, United States
[b] Department of Computer Science, Virginia Tech, United States

ARTICLE INFO

Article history:
Received 16 December 2010
Received in revised form
10 March 2013
Accepted 12 March 2013
Communicated by Zhu Xingquan
Available online 2 May 2013

Keywords:
Feature points
Fisher score
Linear discriminant analysis
Locality

ABSTRACT

Appearance of objects lie in high-dimensional spaces. Feature selection improves not only the efficiency of object recognition but also the recognition accuracy. In this paper, we propose a two-layer learning framework of feature selection using spatial and discriminant influences. The first layer selects a number of feature points of highest integrated influences by integrating spatial and discriminant influences, and the second layer refines the selection in terms of the discriminancy of these feature points measured by orientation histograms of their local appearances. The proposed framework can be categorized as a global appearance based recognition approach. Unlike popular projection methods, such as PCA, LDA, the proposed framework can present visual interpretability of selected features, which is desirable in bioinformatics and medicine informatics. We present two case studies: (i) embryo stage recognition and (ii) face recognition. Our case studies show the effectiveness of the proposed framework.

1. Introduction

Appearances of objects lie in high-dimensional spaces. For a given recognition task, feature selection aims to select most effective features (specifically, feature points) in order to reduce the computational cost of recognition and improve recognition accuracies. Features can be selected based on their spatial influences [15,26,31,53], i.e., the bottom-up scheme [29,28,48,59]. For example, the Harris detector [15] uses gradient auto-correlation of image points to define their spatial influences. The bottom-up scheme aims to output feature points repeatable across different imaging conditions, which helps construct robust and compact representation of image data. The bottom-up scheme has a wide range of applications, such as object recognition [18], image retrieval [33]. The bottom-up feature selection is an important step to build a generative model for object recognition [54,8]. A generative model is basically a graph model with a relatively small number of features [18]. Generative models are strong in addressing "weak-alignment" recognition tasks where the shapes of different objects contain significant variations. Features can also be selected in terms of class or context information, i.e., top-down schemes[13,29,28,48,59]. Gao and Vasconcelos [13] argued that spatial information (such as edge, corners) may not always reveal good saliency of visual objects, and thus proposed a discriminant top-down selection method for visual recognition, where the discriminancy is determined by the maximum marginal diversity [49].

Recently, the integration of the bottom-up and top-down feature selection received extensive attention in the area of visual classification [29,28,48,59], including object detection [34], object recognition [18], and scene understanding [48]. For example, Holub and Perona [18] proposed a model to combine the generative model and Fisher kernels, which brings considerable improvement of the performance of generative models. To speedup object detection, Navalpakkam and Itti [34] proposed a model to integrate bottom-up and top-down attention, where the top-down component uses accumulated statistical knowledge of the visual features of the desired search target and background clutter, to optimally tune the bottom-up maps such that target detection speed is maximized. More related work will be presented in Section 2.

In this paper, we propose a two-layer learning framework for appearance based recognition via a hierarchical usage of spatial and discriminant influences. The proposed framework assume that images (objects) are aligned so that image points at the same location in different images have "correspondence", e.g., their intensities tend to be correlated to each other. In other words, the proposed framework stands on the techniques of image registration [4,17,60], object localization [3,52,11], and image segmentation [55,30,35].

The main idea of the proposed framework is illustrated in Fig. 1. Given a set of training images, the first layer aims to select a

* Corresponding author. Tel.: +1 270 7456225.
E-mail addresses: qi.li@wku.edu (Q. Li), ctlu@vt.edu (C.-T. Lu).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.03.015

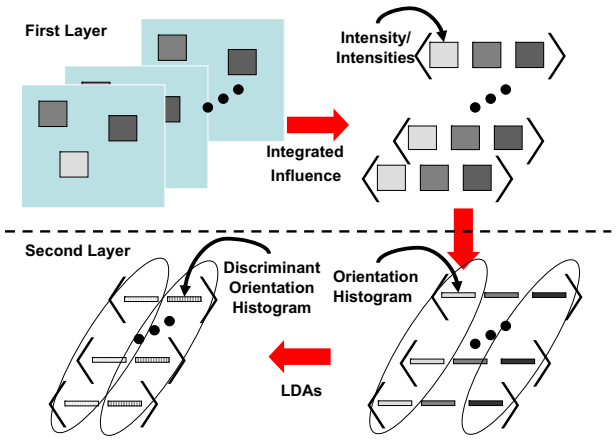

Fig. 1. Two-layer learning framework for appearance based recognition. The first layer selects a number of feature points of maximal integrated influences, and the second layer applies Linear Discriminant Analysis to an ensemble of descriptors of these feature points (constructed by orientation histogram) and obtains an ensemble of most discriminant representations of orientation histograms.

number of feature points (visualized as small blocks in Fig. 1) by integrating spatial and discriminant influences. The major output of the first layer is the locations of feature points. In the second layer, we first construct the orientation histogram (visualized by a row vector in Fig. 1) for the local appearance of each feature point (of each image). The collection of orientation histograms of the same feature points of all training images forms an ensemble of instances (visualized by an ellipse). Linear Discriminant Analysis (LDA) [9,45,1] applies to an ensemble of instances to estimate the discriminancy and compute the discriminant representation of orientation histograms. So the major output of the second layer of the proposed framework is a reduced set of feature points and LDA subspaces associated with the highest discriminancy score.

The rationale of the hierarchical design of the proposed framework lies in the following comparison between intensity blocks and orientation histograms:

- *Efficiency*. Constructing orientation histograms is much more computationally expensive than constructing intensity blocks, which is the rationale of introducing orientation histogram in the second layer rather than the first layer, i.e., being applied to selected feature points rather than all image points.
- *Sensitivity to localization*. Orientation histograms are less sensitive to localization error than intensity blocks, which is the rationale of introducing Linear Discriminant Analysis (LDA) to estimate the discriminancy of orientation histograms. Note that LDA assumes that data fits Gaussian distribution.

In experiments, we present two case studies to demonstrate the effectiveness of the proposed framework. The first case study is on the recognition of developmental stages of Drosophila embryos based on gene expression pattern images [22,14]. The role of Drosophila (fruit fly) in explicating the function and interconnection of animal genes has established the species as a major model organism [44]. In situ hybridization is a recent technique to document gene expression pattern of embryos along their different developmental periods [7]. (An expression region indicates the response of a gene to a probe RNA.) These documents, a set of embryos images contain rich information on the spatio-temporal patterns that are extremely valuable for the study of gene–gene interaction. Dark regions in an embryo image indicate expression patterns of genes. Given two standardized images of embryos (of pixel-to-pixel correspondence) at the same developmental stage, the interaction strength of two genes can be quantified by computing the similarity of expression patterns [23,20,19,38,27,10], e.g., the ratio of overlapping expression regions of the images. Compared with in situ hybridization, the widely used microarray technique reveals very limited spatial pattern information. The gene expressions obtained from microarray images are, precisely speaking, the average expression levels. The second case study is on face recognition where we use dataset PIE [43] that has 68 human faces. These studies convince us the effectiveness of (i) the integrated influences in selecting good feature points, and (ii) the discriminant representation of orientation histograms.

The rest of the paper is organized as follows: In Section 2, we present related work. In Section 3, we introduce locality oriented Fisher discriminant scores. In Section 4, we propose an integrated model. In Section 5, we propose discriminant representation of orientation histograms. Two case studies are presented in Section 6, and conclusions and future work are given in Section 7.

2. Related work

Appearance based recognition can be roughly categorized into two different strategies: (i) global approaches and (ii) local approaches [40]. Global appearance based recognition usually assumes object regions have been aligned. It has been successfully applied to some weakly textured images, such as face images [56]. A global approach is popularly performed in terms of a projection method, such as Principal Component Analysis (PCA) [21], Linear Discriminant Analysis [1], and their high-dimensional variations: Generalized PCA [51], tensor Discriminant Analysis [46], etc. A global appearance can be effectively "encoded" into a very low dimensional vector for recognition, and thus brings appealing recognition efficiency. Besides the advantage of recognition efficiency, projection methods can achieve satisfying recognition accuracy if global appearances do not contain significant local outlier appearances. A limitation of projection methods is that their performance is difficult to interpret. Note that interpretability is desirable in many applications, such as object categorization [39] and bioinformatics [16]. It is worth noting that sparse learning, as a projection based feature extraction scheme, recently received attention [58,24,47,57] since features extracted by a sparse learning method can be interpreted psychologically and physiologically [58].

In contrast to global approaches, local approaches are robust with respect to localization error and local outlier appearances [41,26]. (Precisely speaking, local approaches do not require object localization.) In a local approach, a set of repeatable/stable image point/regions are first extracted by an interest point/region detector [42,26,31], and then distinctive descriptors are constructed to represent an image. However, local approaches are computationally expensive. Moreover, local approaches are not effective for weakly textured objects, such as face images, feature selection method instead of a projection method.

Walther et al. [53] proposed a bottom-up model for selective attention, where bottom-up saliency map is contributed by the color feature maps, intensity feature maps, and orientation feature maps. They showed that the proposed bottom-up visual attention can strongly improve learning and recognition performance in the presence of large amounts of clutter.

Vasconcelos [49] proposed a discriminant feature selection via maximization of marginal diversity (MMD); for multi-class problems, one-versus-all strategy is applied. Vasconcelos and Vasconcelos [50] proposed an information theoretic feature selection to achieve a good balance between maximizing the discriminant power of selected (local) features and minimizing their redundancy. The method is tested on image retrieval, where the

comparison between two images is achieved by the comparison of Gaussian mixtures of the compact sets of discriminant (local) features detected from the images. Gao and Vasconcelos [13] presented a discriminant saliency method, based on MMD [49], to detect visual objects from cluttered backgrounds.

Navalpakkam and Itti [34] proposed a SNR based model to integrate bottom-up and top-down attention for optimizing detection speed, where SNR (signal-noise-ratio) characterizes the discriminant ratio of the spatial influence of target objects over the spatial influence. Navalpakkam and Itti showed the model, with little computational cost in the form of multiplicative top-down gains on bottom-up saliency maps, predicts many reported bottom-up or top-down influences on human visual search behavior.

Holub and Perona [18] proposed a model to combine generative model and Fisher kernels for object recognition. The generative model used in [18] is a constellation model that aims to find optimal appearance and shape parameters $\{\theta_a, \theta_s\}$ during the mapping of interest points to model parts. A Fisher kernel is a gram matrix constructed by "Fisher score" feature that is the derivative of log likelihood of the parameters of a generative model. (Note that Fisher score used in [18] is different from Fisher criterion score used in this paper.)

Zhu et al. [59] formulated bottom-up models as data-driven methods such as Hough transforms and data clustering, and top-down models as templates of objects (targeted in a specific application). Zhu et al. proposed Data Driven Markov Chain Monte Carlo (DDMCMC) to integrate the bottom-up models and top-down models. Mancas [29,28] proposed a bottom-up model based on structures rarity within an image and a top-down model based a mouse-tracking device that builds models of a global behavior for a given kind of image. Toyoda et al. [48] proposed a framework that integrates bottom-up information and top-down information for scene understanding, such as road image labeling. In their framework, bottom-up information is derived from local features of texture and color, and top-down information is generated from a holistic image context.

Fisher criterion score [2] has been widely used for feature selection. In [12], Fisher criterion score is used to select most discriminant features of microarray expression data and achieved substantial improvement of recognition accuracy.

3. Locality oriented Fisher score

Denote p as an image point, c is a class label, J a set of training data, and J_c is a set of training instances in class c, i.e., $J = \cup_c J_c$.

Fisher score was proposed to maximize the ratio of between-class variation over within-class variation. More specifically, given an attribute p, its Fisher criterion score is defined as follows:

$$\begin{aligned}\text{score}(p) &= \text{score}(\{v_j(p)\}_{j\in J}) \\ &= \frac{\sum_c |v_c(p) - v_t(p)|^2}{\sum_c \sum_{j\in J_c} |v_j(p) - v_c(p)|^2},\end{aligned} \tag{1}$$

where v_j is j-th training instance, $v_c(p) = (1/|J_c|)\sum_{j\in J_c} v_j(p)$, and $v_t(p) = (1/|J|)\sum_{j\in J} v_j(p)$. The most discriminant attribute is assigned by the highest Fisher score. Thus by sorting attributes according to their Fisher scores, a number of most discriminant attributes contribute a good feature vector for recognition, e.g., the use of nearest neighbor under Euclidean distance as a classifier. The number of most discriminant attributes is usually determined via cross-validation.

We introduce locality oriented Fisher scores to estimate discriminant influences where the locality is captured by wavelets. Engaging locality in Fisher score evaluation aims to stabilize discriminant features with respect to image noise and illumination conditions. In this paper, we apply one-level wavelet transformation to capture the locality of an image point. With one-level wavelet transformation, an image is decomposed into 4 sub-bands: LL, LH, HL, and HH. Denote $u_j(\cdot; band)$ is a wavelet sub-band of j-the training instance $v_j(\cdot)$, where band is LL, LH, HL, or HH. We propose the following Fisher scores:

$$\begin{aligned}D(p; band) &= D(\{u_j(p; band)\}_{j\in J}) \\ &= \frac{\sum_c |u_c(p; band) - u_t(p; band)|^l}{\sum_c \sum_{j\in J_c} |u_j(p; band) - u_c(p; band)|^l},\end{aligned} \tag{2}$$

where l is a positive number, $u_c(p; band) = (1/|J_c|)\sum_{j\in J_c} u_j(p; band)$, and $u_t(p; band) = (1/|J|)\sum_{j\in J} u_j(p; band)$.

Next, we briefly illustrate the theory why a wavelet can be used to capture the locality. It is known that wavelets have several desirable properties: compact supports, symmetry, and/or high-vanishing moments, orthogonality, etc. Given a wavelet ψ (for simplicity, let us assume it is on $\mathbb{R}$), *compact support* indicates $\psi(x) \equiv 0$ out of some finite interval; *symmetry* indicates $\psi(x_0 - x) = \psi(x)$, for some $x_0 \in \mathbb{R}$; *vanishing moment* k indicates $\int x^l \psi(x)\, dx$, $l = 0, \ldots, k$; *orthogonality* indicates $\int \psi(x)\psi(x-j)\, dx = 0$, $\forall j \in \mathbb{Z}$. Compact support is the key property for a wavelet technique to perform the local analysis. Vanishing moment is also a useful property for local analysis. Note that if a local region is smooth, it can be approximated by some low-order polynomials. Convolving with a wavelet of some-degree vanishing moment, its associated wavelet coefficients are small. Thus the magnitude of wavelet coefficients can characterize the smoothness of a local region. The work in signal or visual processing has found the importance of symmetry. Orthogonality may be arguable depending on what space the data lies in. If the data is in L^2, it is desirable; Otherwise, it may be worthless.

We will use least asymmetric Daubechies wavelet to capture the locality in determining the Fisher criterion score. The least asymmetric Daubechies wavelet is constructed by constraining the phase of the so-called transfer function as close to linear as possible. (More details can be found in [6, Chapter 8].)

It is worth noting that in our Fisher score formulation, we introduce the norm parameter l. In standard Fisher score, l is always fixed as 2, i.e., Euclidean norm. It is known that in resisting outlier attributes, Euclidean norm may not perform best. In the later case study, we will observe the value of this generalization.

4. Integrating spatial and discriminant influences

Recall that p is an image point and J is a set of training data. Denote i the index of a certain spatial filter such as Gradient auto-correlation [15,42], Laplacian [31], and DoG [26]. Denote $\{S_j^i\}_{j\in J}$ as the spatial influence maps of all training images associated with a certain spatial filter. Denote T^0 as the unsupervised operator $T^0(\{S_j^i(p)\}_{j\in J}) = (1/|J|)\sum_{j\in J} S_j^i(p)$, which gives bottom-up feature selection. Denote $u_j^i(\cdot; band)$ is a one-level wavelet sub-band of $S_j^i(\cdot)$. For convenience, we index LL, LH, HL, and HH as 1, 2, 3, and 4 respectively. Denote $T^k(\{S_j^i(p)\}_{j\in J}) = D(\{u_j^i(p; k)\}_{j\in J})$, $k = 1, \ldots, 4$, (see Eq. (2)), as a supervised operator, which gives top-down feature selection.

Our model integrates a set of unsupervised and supervised operators that are applied to a set of spatial influence maps as follows:

$$\text{influence}(p) = \sum_{0\le k\le 4, i} \alpha_{k,i} T^k(\{S_j^i(p)\}_{j\in J}), \quad \text{subject to} \sum_{0\le k\le 4, i} \alpha_{k,i} = 1, \tag{3}$$

where the weight parameters $\alpha_{k,i}$ reveal the prior of different bottom-up and top-down influences in a specific appearance

based recognition task. The weight parameters can be learned by applying cross-validation to training data, i.e., optimal weights are decided by the recognition accuracy on validation data.

After each image point is assigned with a certain influence value, best features can be selected according to the order of their influence. Fig. 2 shows the influence maps of embryo images overlaid by fifty best feature points (i.e., pixels of strongest influence), illustrating the integration of two popular spatial influences—gradient auto-correlation and Laplacian—with discriminant influence, respectively. We can observe that feature points under gradient auto-correlation influence spread in the entire image plane with any specific concentration, and feature points under integrated influence have better concentration. (The higher recognition accuracy achieved by integrated features, shown in later experiments, explains the value of the concentration.) Face images may give us better visual verification on the value on integrated influence. Fig. 3 influences maps and fifty best pixels of face images (from CMU-PIE dataset [43]), from which we can see that most features points under integrated influence occur in the facial areas, such as the eyes, nose and mouth.

With a set of feature points P, we can construct feature vectors (compact image representations) for appearance based recognition. A convenient and efficient way for constructing feature vectors is to use the intensities of those feature points, i.e., $\{I(p)\}_{p\in P}$.

5. Discriminant representations of orientation histograms

This section includes three parts. The first part describes how to construct orientation histograms, the second part presents LDA discriminant orientation histograms, and the third part presents the recognition scheme based on the LDA discriminant orientation histograms.

5.1. Orientation histograms

An orientation histogram is more precisely called *histogram of oriented gradients* [5]. Denote $O(p)$ a neighborhood of an image point p, $q\in O(p)$ a neighboring point of p, $g_q=(I_x(q),I_y(q))$ the gradient of a point q, $\theta(g_q)$ the orientation of the gradient g_q, and

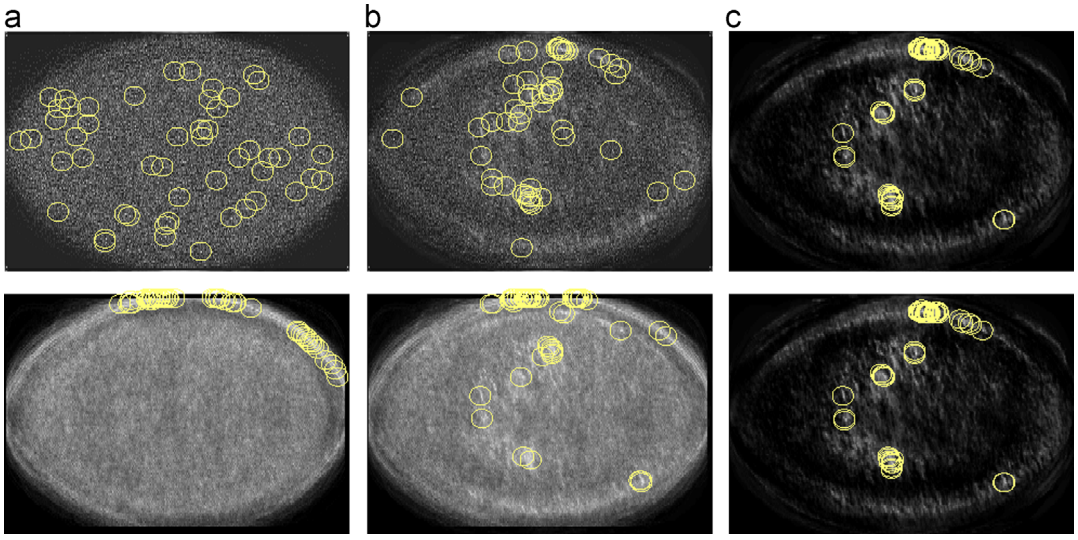

Fig. 2. Influence maps of embryo images overlaid by 50 best feature points. First row = gradient auto-correlation; second row = Laplacian. (a) Spatial, (b) integrated ($\alpha=0.5$), and (c) discriminant.

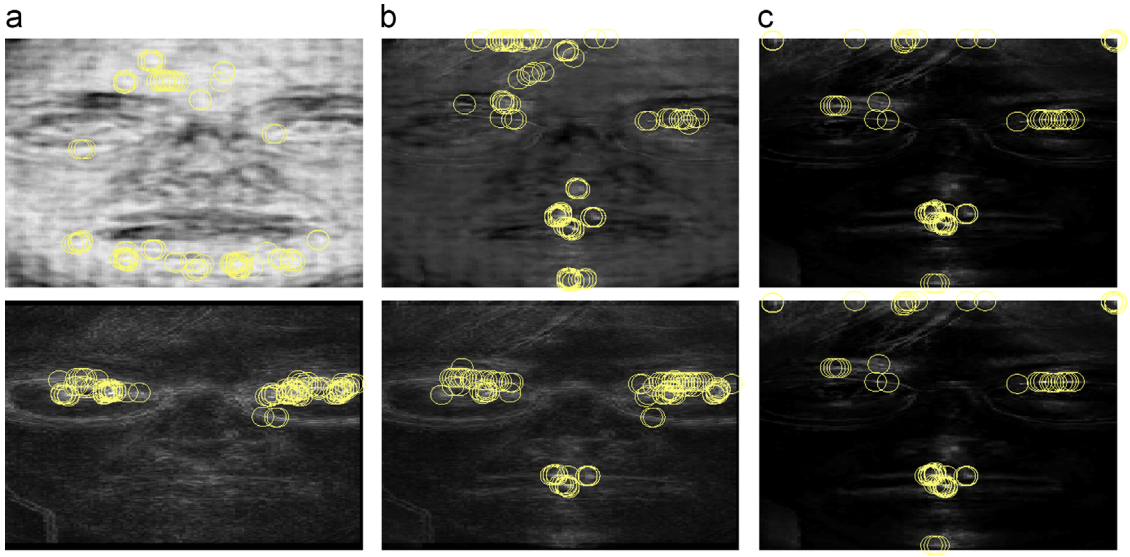

Fig. 3. Influence maps of face images overlaid by 50 best feature points. First row = gradient auto-correlation; second row = Laplacian. (a) Spatial, (b) integrated ($\alpha=0.5$), and (c) discriminant.

$\theta_k, k=1,\ldots,n$ are digitized orientations. An orientation histogram of an image point p is n number of bins, each of which accumulates the magnitudes of the gradients of neighboring points that have the same orientation, this is,

$$h(\theta_k) = \sum_{q\in O(p),\theta(g_q)=\theta_i} \|g_q\|, \quad k=1,2,\ldots,n, \tag{4}$$

where $\|g_q\|$ denotes the magnitude of the gradient g_q. Weighting scheme, e.g., Gaussian weighting can be applied to the magnitude of a gradient g_q based on the distance of q to p.

The number of bins n is an important parameter that may affect the effectiveness of orientation histograms. Theoretically, a larger n (i.e., higher dimension of a histogram) is expected to construct a more distinctive representation (descriptor) of an image point. However, a larger n, in practice, also likely implies an over-refined angular space of gradients, which tends to cause higher sensitivity to imaging conditions (such as illuminations, rotations) and localization errors.

Aiming to construct distinctive representation robust to various imaging conditions, Lowe [26] proposed to a concatenation strategy that concatenates multiple low-dimensional histograms to build a relatively high-dimensional histograms. More specifically, a neighborhood of an image point p will be first sub-divided into $m\times m$ (e.g., $m=4$) of blocks. For each block $O_{ij}(p)$, we build a low-dimensional histogram h_{ij} (e.g., the dimension $n=8$). The final representation is then defined as

$$h=(h_{11},h_{12},\ldots,h_{mm}). \tag{5}$$

It is easy to see that the concatenation of multiple histograms is more robust to localization error [26].

5.2. LDA discriminant orientation histograms

Based on orientation histograms (of a set of feature points), we next apply Linear Discriminant Analysis (LDA) [9,1] to rank the discriminancy of feature points and discriminant representation.

Different from the single-attribute strategy used in Fisher criterion score above, the Fisher criterion for LDA concerns with the discriminant ratio covering all attributes. More specifically, for an ensemble of orientation histograms, the Fisher criterion for LDA computes an optimal discriminant linear projection W as follows:

$$W=\arg\max_W \frac{|W^T S_b W|}{|W^T S_w W|}, \tag{6}$$

where S_b and S_w are so-called between-class and within-class scatter matrices, respectively, and W^T indicates the transpose of W. More specifically, given a training set H of N orientation histograms (see Eq. (5)), S_w and S_b are constructed as follows:

$$S_w=\frac{1}{N}\sum_{c=1}^{L}\sum_{h\in H_c}(h-h_c)(h-h_c)^T$$

$$S_b=\frac{1}{N}\sum_{c=1}^{L}N_c(h_c-h_t)(h_c-h_t)^T$$

L is the number of classes, H_c is the collection of orientation histograms in c-th class, N_c is the size of H_c, h_c is the mean of H_c, and h_t is the (total) mean of H.

Note that $|W^T S_b W|/|W^T S_w W|$, as the discriminancy, will be used to rank the corresponding feature point.

The optimal W can be computed by the generalized eigen-analysis, i.e., $S_w^{-1}S_b$. The rank of W is usually chosen as $k-1$ (k is the number of classes), which implies effective dimensionality reduction in many real-life applications, in contrast to Fisher feature selection. LDA can also be used to visualize the degree of linear separability of high-dimensional data.

LDA has the singularity issue for small sample size problems, i.e., the dimension is larger than the number of training samples. A popular method to overcome the singularity issue is to apply PCA (Principal component analysis) to reduce the dimension of original data before LDA (so called Fisherface/PCA+LDA [1]). Under the basic constraint that the retained dimension should be less than the number of training samples, it is important to choose an optimal retained dimension, which can be obtained via cross-validation.

5.3. Ensemble recognition

As illustrated in Fig. 1, the appearance based recognition under the proposed framework should be, more precisely speaking, called ensemble recognition since an image is represented by multiple instances (discriminant orientation histograms) associated with different point location. More specifically, given a set of training (gallery) images, we perform the following training steps:

(1) Apply the first layer of the proposed framework to select a number of feature points.
(2) Construct orientation histograms for each selected feature point.
(3) Compute LDA discriminancy and LDA subspaces W.
(4) Select d most discriminant feature points.
(5) Compute d LDA discriminant representations.

Given a query image, we perform the following recognition steps:

(1) Construct d orientation histograms based on d most discriminant feature points.
(2) Compute d LDA discriminant representations.
(3) Apply a nearest neighbor classifier to each representation.
(4) Apply majority vote to the output of step (3).

6. Case studies

In this section, we test the proposed framework in two case studies: (i) recognition of stages of Drosophila embryos [23], and (ii) face recognition. For each case study, images have been aligned, which is an assumption for the application of the proposed framework. An embryo is aligned based on the centroid and the orientation of the contour of the targeting embryo object, and a face image is aligned based on the locations of the two eyes' centers of a face object.

It is worth noting that the studies of localization techniques provide a base for the above assumption. Automatic localization of Drosophila embryos recently received intensive attentions [37,36,10,27,38,25], and the-state-of-the-art techniques achieved successful rates higher than 92% [25]. Face localization/detection received longer attentions, and some proposed techniques, such as Viola–Jones method [52], have been widely used in real-life applications and research.

The datasets used for our case studies are the following:

- *Embryo*. Our dataset has 500 images of fruit fly embryo, in three classes. The goal of classifying embryo images is to identify embryo developmental stages that is an important step towards gene expression analysis. The raw images, as shown in Fig. 4, contain severe illumination variations. We apply the histogram equalization method to normalize embryo images. Recall that the nature of weak texture of embryo images

Fig. 4. Weakly textured appearances of fruit fly embryos in three different stages. Images in the same row belong to the same stage. Embryos in an earlier stage have smoother contours and simpler appearance textures. Dark regions are gene expression regions. Gene expression regions may spread in the entire appearance of an embryo (as shown in the top-left image), or they may be just a small portion of the appearance of an embryo (as shown in the bottom-middle image). The variation of gene expressions is one of challenges in the recognition of embryo stages.

motivates us to explore the opportunity of using relatively large number of features.

- *CMU-PIE* [43]. PIE has 68 human faces, each of which has 22 illumination instances. We also apply histogram equalization to normalize the face images.

In our experiments, a dataset is randomly split into two: one half is used as training and validation set, and the other half as the test set. We run 5-fold on training and validation data to decide the optimal parameters: weights (bottom-up and top-down priors) and the number of feature points. To reduce the variability, the splitting is repeated 5 times and the resulting accuracies are averaged. The number of feature points (n) in our experiment is from 400 to 2000. We use nearest neighbor as the classifier. Our test configuration consists of a computer of CPU Pentium 4 (3.40 GHz) and Memory 4 G with Matlab.

6.1. Fisher score: standard versus locality oriented

First, we present a comparison between the standard Fisher score and the locality oriented Fisher score ($l=1$ or 2) in two different appearance based recognition tasks. Table 1 shows the results, and it is clear that the locality oriented Fisher score outperforms the standard Fisher score, and two Fisher scores are comparable to each other. We observe that the performance of norm $l=2$ is slightly better than norm $l=1$, in the case of pure discriminant selection. However, as we will see soon, the observation will be different when the locality oriented Fisher scores are integrated with a certain bottom-up scheme, which in turn leads to the use of both norm in the integrated model.

Furthermore, we measure the performance of locality oriented discriminant influence with different norms integrated with a certain bottom-up scheme. Fig. 5 illustrates the behavior under a simple version of integrated model (spatial influence is contributed by gradient auto-correlation only). We can observe that highest accuracy is achieved by the integrated influence associated with norm $l=1$. It is worth noting that this interesting observation occurs consistently across varied n, which reveals the benefit of introducing l in the locality oriented Fisher criterion score. In the later experiments, we use two discriminant operators, i.e., T^1 and T^2 are associated with norm 1 and 2, respectively.

Fig. 6 shows the validation accuracy in cross-validation, where X-axis indicates the weight α, Y-axis indicates the length of feature vectors, and Z-axis indicates the validation accuracy. Fig. 6(a) and (b) are associated with gradient auto-correlation, and Laplacian (two spatial influence assignments), and the norm l in the discriminant influence is 2. First of all, Fig. 6 gives an example that discriminant influence does not always outperform spatial influence. More importantly, Fig. 6 shows the mutual benefit of spatial and discriminant influences, for example, the highest accuracy is always achieved by a certain degree of integration of spatial and discriminant influence. The optimal parameters for gradient auto-correlation are ($\alpha=0.5$, $n=400$), and the ones for Laplacian are ($\alpha=0.6$, $n=2000$).

Table 1
Recognition accuracy. A comparison between standard Fisher score and locality oriented Fisher scores.

Discriminant methods	Embryo	Face (PIE)
Standard Fisher score	0.80	0.95
LO Fisher score ($l=1$)	0.83	0.96
LO Fisher score ($l=2$)	0.84	0.97

In the following, we have a visual comparison among the linear separability of these feature vectors where the feature points are selected via spatial, discriminant and integrated influence, respectively. (Note that linear separability is desirable to support efficient classifiers.) We use embryo images as examples, and apply Linear Discriminant Analysis (LDA) to visualize the feature vectors in 2-D plane. The dimension of embryo images is 320×128. Our data contains three classes (leading to two-dimensional LDA space). We will show a PCA+LDA representation as a comparison. Fig. 7 shows four different LDA representation. The first two classes of embryo data are shown for the clarity of comparison of the representation. The bold labels indicate the data items violating linear separability. From Fig. 7, we can see that the integrated influence contributes to feature vectors of best linear separability. This example gives us an insight of the effectiveness of integrating spatial and discriminant influences in improving the linear separability of the image representation.

6.2. Main results

Next, we test the performances of various features by the two case studies: (i) embryo stage recognition and (ii) face recognition.

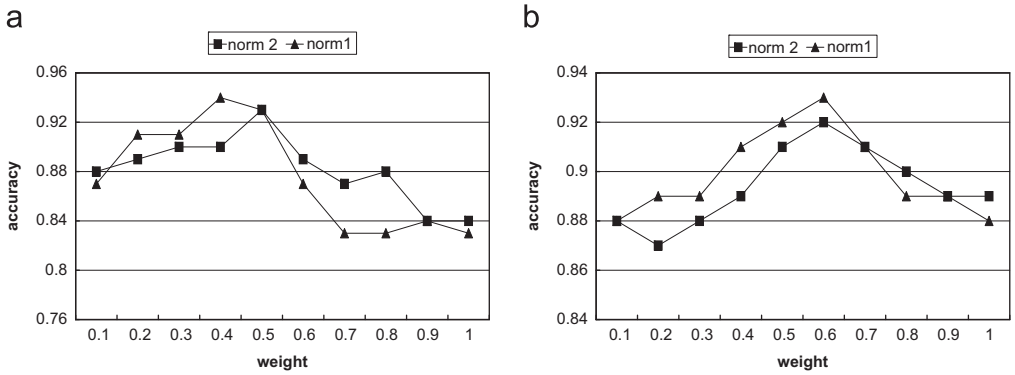

Fig. 5. A comparison between norm $l=2$ and $l=1$. The highest accuracy is achieved by the integrated influence associated with norm 1. (a) 400 features and (b) 2000 features.

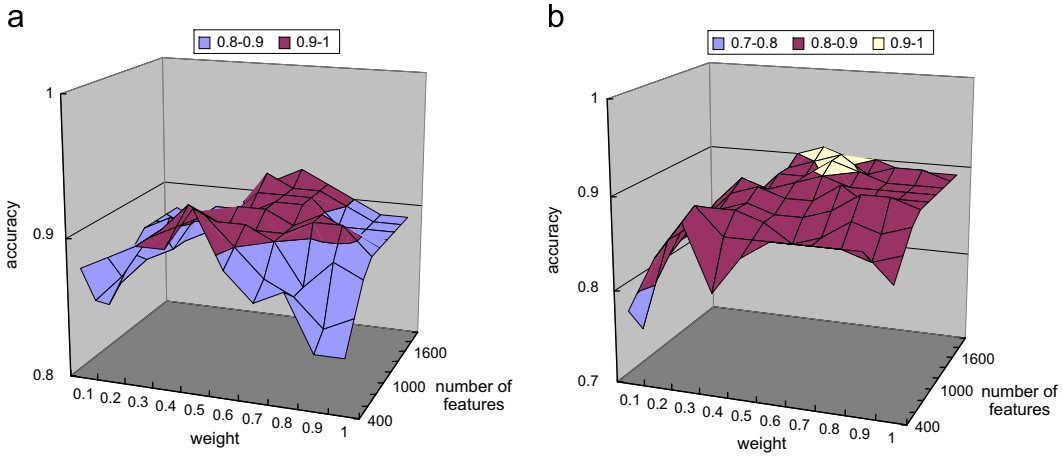

Fig. 6. Integrated influence with norm $l=2$ in discriminant influence. The optimal parameters for gradient auto-correlation are $\alpha=0.5$, $n=400$, and the ones for Laplacian are $\alpha=0.6$, $n=2000$. (a) Gradient auto-correlation and (b) Laplacian.

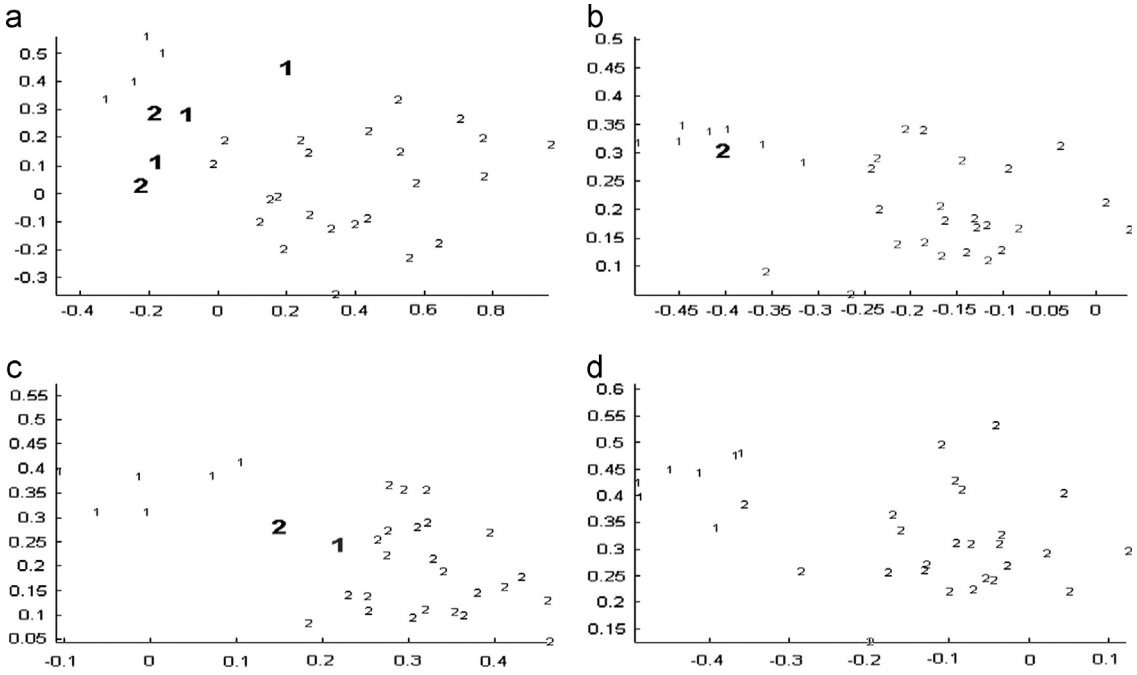

Fig. 7. Visualization of 4 different feature vectors. The bold labels indicate the data items violating linear separability. The integrated influence contributes feature vectors of best linear separability. (a) PCA, (b) spatial influence, (c) discriminant influence, and (d) integrated influence.

Tables 2 and 3 show the comparison of the effectiveness of one-layer features, simple discriminant representation of one-layer features, and two-layer features under different integration of spatial influence. Specifically, the second and third column of the tables show recognition accuracies under the gradient auto-correlation and Laplacian spatial influence, respectively. The last column shows the accuracy under both spatial influence.

The first rows of Table 2 (on embryo stage recognition) and Table 3 (on face recognition) show the test results using features generated by the first layer of the proposed framework. Recall that an image, in terms of the first layer, is represented as a single vector constructed by the intensities of selected feature points. The results convince the effectiveness of the integration model in integrating different spatial influence maps, i.e., higher accuracy and smaller deviation. It is also worth noting that the integrated model outperforms some baseline methods. For example, on face recognition, using entire face images as representation gives us fairly low recognition accuracy (0.64). With Fisherface technique [1], the accuracy is around 0.90.

The second rows of the tables show the results using LDA representations of first-layer feature vectors, denoted as *1st layer + LDA*. As well as the test above, an image under this test is represented a single vector too. The results in Tables 2 and 3 show that LDA representation degrades the performance of the feature vectors. After all, the dimension of the LDA representation is much lower than the dimension of the feature vectors.

The third rows of the tables show the results using two-layer features. Unlike the first two rows, an image under this test is represented multiple vectors, i.e., a number of discriminant representations of orientation histograms. We apply exhaustive search scheme to the training data to select an optimal number of representations from a range of values from 10 to 100, and obtain that optimal numbers for two case studies are 60 and 85, respectively. Recall that the recognition output is based on the majority vote on the nearest neighbor classifier applied to different ensembles of discriminant representations. The results in Tables 2 and 3 demonstrate the effectiveness of the proposed framework—it not only outperforms the 1st layer+LDA significantly, but also improves the performance of the first layer.

Table 4 shows a comparison of the first layer, the second layer, and two layers in terms of recognition accuracy and recognition time (measured in second). We can observe that the first and second layers are competitive to each other in terms of the recognition accuracy, while the second layer has much higher computational cost than the first layer. The computational cost of the two-layer approach is much lower than the second layer's. But it is worth noting that the training cost is very high—the training time in two case studies is 1830 and 6250 seconds, respectively.

Table 5 shows a comparison of the proposed framework with four appearance-based recognition approaches. Two of them are global approaches: (i) PCA [21] and (ii) LDA [1]), and another two are local approaches: (i) SIFT [26] and (ii) MSER [32]. We can observe that the proposed framework outperforms the four existing approaches. We can also observe the superiority of global approaches over local approaches in these two case studies. Note that both embryo images and face images are weakly textured, which is a challenge for local approaches.

Table 2
Recognition accuracy (with deviation) on embryo stage recognition. The second and third column of the tables show recognition accuracies under the gradient auto-correlation and Laplacian spatial influence, respectively. The last column shows the accuracy under both spatial influence.

Methods	Gradient auto-correlation	Laplacian	Both
1st layer	0.93 (0.04)	0.91 (0.05)	0.94 (0.03)
1st layer+LDA	0.83 (0.04)	0.82 (0.05)	0.85 (0.03)
Two layers	0.94 (0.03)	0.93 (0.04)	0.95 (0.03)

Table 3
Recognition accuracy (with deviation) on face recognition. The second and third column of the tables show recognition accuracies under the gradient auto-correlation and Laplacian spatial influence, respectively. The last column shows the accuracy under both spatial influence. The accuracy achieved by Fisherface is around 90%.

Methods	Gradient auto-correction	Laplacian	Both
1st layer	0.98 (0.01)	0.99 (0.01)	1.00 (0.0)
1st layer +LDA	0.92 (0.02)	0.92 (0.02)	0.94 (0.01)
Two layers	0.99 (0.01)	0.99 (0.01)	1.00 (0.0)

Table 4
Comparison of the first layer and the second layer in terms of recognition accuracy and recognition time (in second). Without feature selection, the computational cost of the second layer is much higher than the first layer.

Methods	Embryo		Face	
	Accuracy	Time	Accuracy	Time
1st layer	0.94	1	1.0	2
2nd layer	0.95	12	1.0	23
Two layer	0.95	3	1.0	5

Table 5
Comparison of the recognition accuracy of the proposed framework with four appearance-based recognition approaches. Results show the superiority of the proposed framework.

Dataset	Proposed	Global		Local	
		PCA	LDA	SIFT	MSER
Embryo	0.95	0.78	0.71	0.61	0.63
Face	1.00	0.85	0.93	0.89	0.91

7. Conclusions and future work

In this paper, we propose a two-layer framework for appearance based recognition using spatial and discriminant influence. The hierarchical design of the proposed framework is mainly motivated by the high computational cost of the construction of orientation histograms. We present two case studies to demonstrate the effectiveness of the proposed framework. Note that an assumption of the proposed framework is that images (objects) are aligned. In the future, we plan to integrate shape analysis with the two-layer framework for object recognition.

References

[1] P.N. Belhumeur, J. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. In: European Conference on Computer Vision, vol. 1, 1996, pp. 45–58.
[2] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, UK, 1995.
[3] M. Brejl, M. Sonka, Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples, IEEE Trans. Med. Imaging 19 (10) (2000) 973–985.
[4] L.G. Brown, A survey of image registration techniques, ACM Comput. Surv. 24 (December (4)) (1992) 325–376.
[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 886–893.
[6] I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, 1992.

[7] P. Tomancak, et al., Systematic determination of patterns of gene expression during drosophila embryogenesis, Genome Biol. 3 (12) (2002) 1–14.
[8] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: IEEE Computer Vision and Pattern Recognition, 2003, pp. 264–271.
[9] R.A. Fisher, The use of multiple measurements in taxonomic problems, in: Annals of Eugenics, vol. 7, 1936, pp. 179–188.
[10] E. Frise, A. S Hammonds, S.E. Celniker, Systematic image-driven analysis of the spatial drosophila embryonic expression landscape, Mol. Syst. Biol. 6 (2010) 345.
[11] B. Fulkerson, A. Vedaldi, S. Soatto, Class segmentation and object localization with superpixel neighborhoods, in: International Conference on Computer Vision, 2009, pp. 670–677.
[12] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, Michèl Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914.
[13] D. Gao, N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes, in: Neural Information Processing Systems (NIPS), Electronic edition, 2004.
[14] R. Gurunathan, B.V. Emden, S. Panchanathan, S. Kumar, Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: binary feature versus invariant moment digital representations, BMC Bioinformatics 5 (2004) 202.
[15] C. Harris, M. Stephens, A combined corner and edge detector, in: Proceedings of the 4th Alvey Vision Conference, Manchester, 1988, pp. 147–151.
[16] A.C. Haury, P. Gestraud, J.P. Vert, The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures, PLoS ONE 6 (12) (2011) e28210.
[17] D.L.G. Hill, P.G. Batchelor, M. Holden, D.J. Hawkes, Medical image registration, Phys. Med. Biol. 46 (1) (2001) 1–45.
[18] A. Holub, M. Welling, P. Perona, Combining generative models and fisher kernels for object recognition, in: IEEE International Conference on Computer Vision, 2005, pp. 136–143.
[19] S. Ji, Y.X. Li, Z.H. Zhou, S. Kumar, J. Ye, A bag-of-words approach for drosophila gene expression pattern annotation, BMC Bioinformatics 10 (2009) 119.
[20] S. Ji, L. Sun, R. Jin, S. Kumar, J. Ye, Automated annotation of drosophila gene expression patterns using a controlled vocabulary, Bioinformatics 24 (17) (2008) 1881–1888.
[21] I.T. Jolliffe, Principal component analysis, J. Educ. Psychol. 24 (1986) 417–441.
[22] S. Kumar, K. Jayaraman, S. Panchanathan, R. Gurunathan, A. Marti-Subirana, S. J. Newfeld, Best: a novel computational approach for comparing gene expression patterns from early stages of drosophila melanogaster development, Genetics 162 (4) (2002) 2037–2047.
[23] S. Kumar, K. Jayaraman, S. Panchanathan, R. Gurunathan, A. Marti-Subirana, S. J. Newfeld, Best: a novel computational approach for comparing gene expression patterns from early stages of drosophila melanogaster development, Genetics 16 (4) (2002) 2037–2047.
[24] J. Li, D. Tao, On preserving original variables in Bayesian PCA with application to image analysis, IEEE Trans. Image Process. 21 (12) (2012) 4830–4843.
[25] Q. Li, C. Kambhamettu, Contour extraction of drosophila embryos, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (6) (2011) 1509–1521.
[26] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[27] D.L. Mace, N. Varnado, W. Zhang, E. Frise, U. Ohler, Extraction and comparison of gene expression patterns from 2d RNA in situ hybridization images, Bioinformatics 15 (26(6)) (2010) 761–769.
[28] M. Mancas, Relative influence of bottom-up and top-down attention, in: 5th International Workshop on Attention in Cognitive Systems, 2008, pp. 212–226.
[29] M. Mancas, C. Mancas-Thillou, B. Gosselin, B.M. Macq, A rarity-based visual attention map-application to texture description, in: IEEE International Conference on Image Processing, 2006, pp. 445–448.
[30] A.M. Mharib, A.R. Ramli, S. Mashohor, R.B. Mahmood, Survey on liver CT image segmentation methods, Artif. Intell. Rev. 37 (2) (2012) 83–95.
[31] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, Int. J. Comput. Vis. 60 (1) (2004) 63–86.
[32] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.
[33] K. Mikolajczyk, C. Schmid, Indexing based on scale invariant interest points, in: IEEE International Conference on Computer Vision, vol. 1, Vancouver, Canada, 2001, pp. 525–531.
[34] V. Navalpakkam, L. Itti, An integrated model of top-down and bottom-up attention for optimal object detection, in: Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 1–7.
[35] J.A. Noble, D. Boukerroui, Ultrasound image segmentation: a survey, IEEE Trans. Med. Imaging 25 (8) (2006) 987–1010.
[36] J.Y Pan, A.G.R. Balan, E.P. Xing, A.J.M. Traina, C. Faloutsos. Automatic mining of fruit fly embryo images, in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2006, pp. 693–698.
[37] H. Peng, E.W. Myers, Comparing *n situ m*rna expression patterns of *drosophila* embryos, in: Research in Computational Molecular Biology (RECOMB), 2004, pp. 157–166.
[38] K. Puniyani, C. Faloutsos, E.P. Xing, Spex2: automated concise extraction of spatial gene expression patterns from fly embryo ISH images, Bioinformatics 26 (12) (2010) i47–i56.
[39] A. Rebai, A. Joly, N. Boujemaa, Blasso for object categorization and retrieval: towards interpretable visual models, Pattern Recognition 45 (6) (2012) 2377–2389.
[40] P.M. Roth, M. Winter, Survey of Appearance-Based Methods for Object Recognition, Technical Report ICG-TR-01/08, Institute for Computer Graphics and Vision, Graz University of Technology, Austria, 2008.
[41] C. Schmid, R. Mohr, Local grayvalue invariants for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 19 (May (5)) (1997) 530–534.
[42] C. Schmid, R. Mohr, C. Bauckhage, Evaluation of interest point detectors, Int. J. Comput. Vis. 37 (2) (2000) 151–172.
[43] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression (PIE) database, in: Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition (FG'02), 2002.
[44] P. Simpson, Evolution of development in closely related species of flies and worms, Nat. Rev. Genet. 3 (12) (2002) 907–917.
[45] D.L. Swets, J.J. Weng, Using discriminant eigen features for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell 18 (8) (1996) 831–836.
[46] D. Tao, X. Li, X. Wu, S.J. Maybank, General tensor discriminant analysis and gabor features for gait recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1700–1715.
[47] X. Tian, D. Tao, Y. Rui, Sparse transfer learning for interactive video search reranking, ACM Trans. Multimedia Comput. Commun. Appl. 8 (3) (2012) 26.
[48] T. Toyoda, K. Tagami, O. Hasegawa, Integration of top-down and bottom-up information for image labeling, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1106–1113.
[49] N. Vasconcelos, Feature selection by maximum marginal diversity, in: Neural Information Processing Systems (NIPS), 2002.
[50] N. Vasconcelos and M. Vasconcelos, Scalable discriminant feature selection for image retrieval and recognition, in: Computer Vision and Pattern Recognition, vol. 2, 2004, pp. 770–775.
[51] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1945–1959.
[52] P.A. Viola, M.J. Jones, Robust real-time face detection, in: International Conference on Computer Vision, 2001, p. 747.
[53] D. Walther, U. Rutishauser, C. Koch, P. Perona, Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, Comput. Vis. Image Understanding 100 (2005) 41–63.
[54] M. Weber, Unsupervised Learning of Models for Object Recognition, Ph.D. Thesis, Department of Computational and Neural Systems, Caltech, Pasadena, CA, 2000.
[55] H. Zhang, J.E. Fritts, S.A. Goldman, Image segmentation evaluation: a survey of unsupervised methods, Comput. Vis. Image Understanding 110 (2) (2008) 260–280.
[56] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition in still and video images: a literature survey, ACM Comput. Surv. 35 (4) (2003) 399–458.
[57] T. Zhou, D. Tao, Double shrinking sparse dimension reduction, IEEE Trans. Image Process. 22 (1) (2013) 244–257.
[58] T. Zhou, D. Tao, X. Wu, Manifold elastic net: a unified framework for sparse dimension reduction, Data Min. Knowl. Discov. 22 (3) (2011) 340–371.
[59] S.C. Zhu, R. Zhang, Z. Tu, Integrating bottom-up/top-down for object recognition by data driven Markov chain Monte Carlo, in: IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 1738–1745.
[60] B. Zitova, J. Flusser, Image registration methods: a survey, Image Vis. Comput. 21 (11) (2003) 977–1000.

Qi Li is an Associate Professor of the Department of Computer Science at Western Kentucky University. He received his Ph.D. in Computer Science from University of Delaware in 2006. His current research interest include pattern recognition, computer vision, machine learning, and bioinformatics.

Chang-Tien Lu is an Associate Professor of the Department of Computer Science at Virginia Tech. He received his Ph.D. in Computer Science from University of Minnesota in 2001. His current research interest include data mining, spatial databases, spatial query processing, data warehousing, and geographic information systems.