Multi-scale gist feature manifold for building recognition

Cairong Zhao, Chuancai Liu, Zhihui Lai

Abstract

Multi-scale gist (MS-gist) feature manifold for building recognition is presented in the paper. It is described as a two-stage model. In the first stage, we extract the multi-scale gist features that represent the structural information of the building images. Since the MS-gist features are extrinsically high dimensional and intrinsically low dimensional, in the second stage, an enhanced fuzzy local maximal marginal embedding (EFLMME) algorithm is proposed to project MS-gist feature manifold to low-dimensional subspace. EFLMME aims to preserve local intra-class geometry and maximize local interclass margin separability of MS-gist feature manifold of different classes at the same time. To evaluate the performance of our proposed model, experiments were carried out on the Sheffield buildings database, compared with the existing works: (a) the visual gist based building recognition model (VGBR) and (b) the hierarchical building recognition model (HBR). Moreover, EFLMME is evaluated on Sheffield buildings database compared with some linear dimensionality reduction methods. The results show that the proposed model is superior to other models in practice of building recognition and can handle the building recognition problem caused by rotations, variant lighting conditions and occlusions very well.

Keywords: Multi-scale gist feature, Fuzzy gradual neighbor graph, Building recognition, Manifold learning

1. Introduction

Object recognition is an important area in computer vision research. There is a large amount of literatures [1–6] on general object recognition. As a relatively specific recognition task in object recognition, building recognition has been somewhat neglected. Developing a computational model of building recognition is difficult because building images may be taken from different viewpoints (please see Figs. 3(c) and 4(a)), under different lighting conditions (please see Fig. 4(c), or suffered from occlusions from trees, moving vehicles, other buildings or themselves (see Figs. 3(a) and 4(b)). More details can be seen in the experiment section. This domain becomes more and more interesting to researchers since building recognition can be utilized in many practical application areas, such as architectural design, building labeling in videos, robot vision or localization [7] and mobile device navigation [8,9].

Hutchings and Mayol-Cuevas [9] proposed to use the Harris corner detector [10] based on the local auto-correlation function to extract interest points for matching buildings in the world space for mobile device. In [11], perceptual grouping rules were applied to explore the semantic relationships between different primitive image features. And based on these rules, a Bayesian framework was proposed for content-based image retrieval (CBIR) for buildings. Stassopoulou [12] proposed to use Bayesian Networks for detecting and recognizing buildings from digital orthophotos. The semantic relationships among low-level visual features were explored using the perceptual grouping in a hierarchical way, which was as closely as possible to fit human behavior. Li et al. [13] proposed to use SIFT [14] descriptors to extract the distinctive invariant features from images, which provides reliable matching between different views of objects or scenes. Fritz et al. [15] proposed to apply the ‘Informative Descriptor Approach’ on SIFT features (i-SIFT descriptors) to provide both robust building detection and recognition by extracting color, orientation and spatial information for each line segment. Zhang and Kosecka [16] proposed a hierarchical approach for building recognition based on localized color histograms in the first step and refined matching SIFT descriptors in the second step. With the quick retrieval of a small number of candidate buildings using localized color histograms, this method achieved some improvement in efficiency in building image retrieval. Nevertheless, building recognition systems mentioned above suffer from the detection of low-level vision features [18], e.g., line segments, vanishing points, etc. Since low-level features cannot reveal the truly semantic concepts of images, Li et al. [17],

...
proposed a building recognition model integrating low-level vision features into mid-level features, which attempt to reveal the truly semantic concepts of the images. Due to approximately semantic feature extraction, this method obtains some improvement in building recognition.

By pooling together the activity of local low-level feature detectors across large regions of visual field, Oliva and Schyns [19] and Oliva and Torralba [20] built a holistic and low-dimensional representation of the structure of a scene. High-level properties of an image such as the degree of perspective have been found to be correlated with the configuration of low-level image features [21–23]. Evidence from the psychophysics literature suggest that our visual system analyzes global statistical summary of the image in a pre-selective stage of visual processing or at least, with minimal attentive resources [24]. This suggests that a reliable gist representation of building also can be abstracted from spatial information by performing 2D Fourier transform analyses in image. Series of behavioral experiments [25–27] showed that visual input was processed at different spatial scales (from low to high spatial frequency), which offer different qualities of information for recognition purpose. Additional experiments [28] showed that the visual system could select a certain spatial scales to process depending on task constraints. As a specific recognition task in pattern recognition, the building recognition system may capture the holistic information of building at different scales. Based on these observations, we propose to use multi-scale gist (MS-gist) feature to describe building image, which is the first contribution of this paper.

Since the MS-gist features of a building image taken from different viewpoints or under different lighting conditions are extrinsically high dimensional and intrinsically low dimensional, it is necessary for us to obtain the more compact representation of the gist features by further dimensionality reduction. Researchers have developed many useful dimensionality reduction techniques [8,29–36]. Principal component analysis (PCA) [32] and linear discriminant analysis (LDA) [30,33] are two well-known linear subspace learning methods. PCA projects the data along the direction of the largest variance of the data; while LDA maximizes the ratio of the between-class scatter to the within-class scatter. But both of them do not take the local geometric structure of the feature manifold into account. Recently, manifold learning becomes an important field in pattern recognition and machine learning. Manifold learning algorithms for dimensionality reduction aim at discovering and preserving the intrinsic low-dimensional compactness representation of the high-dimensional data. The most well-known linear dimensionality reduction method based on manifold learning is locality preserving projections (LPP) [34], which learns a linear subspace approximating the eigenfunctions of Laplacian Eigenmap [35]. The geometric structure based these methods frequently use the K-nearest graph to model the local structure of the dataset. The graphs used in the algorithms are very important since each element of the graphs reflects the relationship between two data points. However, a common drawback of existing graph-based manifold learning algorithms is that the graphs are constructed with Gaussian kernel function or binary pattern in local K-nearest neighborhood data points. Since the distances between the samples in local K-nearest neighbors might also vary in a big range, the graphs constructed in this way might have potential disadvantages that the weights are not in accordance with the nature relations of the data in actual applications. Therefore we present a new approach called enhanced fuzzy local maximal marginal embedding (EFLMME). In EFLMME, two novel graphs called enhanced fuzzy neighborhood graphs are constructed by fuzzy k-nearest neighbor (FKNN) method [37,38]. The fuzzy k-nearest neighbor method is utilized to model the nature distribution information of original samples, and this information is utilized to redefine the affinity weights of the intra-class and the interclass neighborhood graphs instead of using the binary pattern or Gaussian kernel function. With the two novel graphs, an enhanced fuzzy local maximal marginal embedding algorithm (EFLMME) is proposed to further reduce the dimensionality of multi-scale gist feature manifold. This is the second contribution of this paper.

Conclusively, in this paper, we mainly focus on the study of effective and robust feature extraction in the building recognition model. Since the features are sampled from a low-dimensional manifold embedded in high-dimensional space [39], it is essential to obtain a suitable mapping to implement the dimensionality reduction. The proposed model is described as a two-stage model. In the first stage, we capture MS-gist feature representation of the structure of the building images. In the second stage, an enhanced fuzzy local maximal marginal embedding algorithm (EFLMME) is proposed to further reduce dimensionality of MS-gist feature manifold.

To evaluate performance of our proposed model, experiments were carried out on Sheffield buildings database, compared with existing works: (a) the visual gist based building recognition model (VGBR): each image is described by a gist feature based on visual features [17]; (b) the hierarchical building recognition model (HBR): each image is represented by a local color histograms indexing vector [16]. At the same time, EFLMME is evaluated on Sheffield buildings database compared with PCA, LDA, LPP and the recently proposed DLA. The results show that the proposed model is an effective method.

The remainder of this paper is organized as follows: Section 2 describes MS-gist feature representation of the building image. Section 3 introduces enhanced fuzzy local maximal marginal embedding algorithm. Section 4 evaluates the recognition performance on Sheffield building database. Section 5 presents the conclusions.

2. MS (multi-scale)-gist feature representation

2.1. Outline of the holistic representation of the spatial envelope

Evidence from the psychophysics literature suggests that human visual system analyzes global statistical summary of the image in a pre-selective stage of visual processing [24]. By pooling together the activity of local low-level feature detectors across large regions of visual field, Oliva and Torralba [20] and Oliva [21] built a holistic and low-dimensional representation of the structure of a scene. Here, it is introduced briefly as follows.

Let \( X \) be a color image with resolution of \( N \times M \), and let its three color components be \( R \), \( G \) and \( B \). The image is converted to one intensity image

\[
i(x,y) = \frac{1}{3}(R+G+B)
\]

(1)

Next, the intensity image was discrete Fourier transformed. It is defined as

\[
I(f_x,f_y) = \sum_{x=-M/2}^{M/2-1} \sum_{y=-M/2}^{M/2-1} i(x,y) h(x,y) e^{-2\pi i (fx x + fy y)}
\]

(2)

where \( f_x \) and \( f_y \) are the spatial frequency variables. \( h(x,y) \) is a circular Hanning window to reduce boundary affects. \( A[f_x,f_y] = |I(f_x,f_y)| \) is the Fourier amplitude spectrum that is a \( N \times (M/2-1) \) element array of real numbers. Clearly, the amplitude spectrum provides high dimensional representations of the input image. Dimensionality reduction is achieved by the principal component analysis (PCA) by considering only the KL functions that account for the maximal variability of the signal described. Therefore, the
coefficients of decompositions are obtained as the holistic and low-resolution representation of the input image.

2.2. The MS-gist feature representation

Series of behavioral experiments [25–27] showed that visual input was processed at different spatial scales. Different spatial scales offer different qualities of information for recognition purpose. As a specific recognition task in pattern recognition, the holistic information of building images at different scales may be helpful to improve building recognition. Based on this observation, we propose a MS-gist representation for building recognition. The procedure is described in detail as follows.

Firstly, the color building image with resolution of \( N \times M \) is converted to one intensity image. Then we decompose the intensity image at multiple spatial scales

\[
p_b(x,y) = \tilde{h}(x,y)
p_s(x,y) = (\downarrow 2)p_{s-1}(x,y), s = 0, \ldots, S
\]

where \( \downarrow 2 \) refers to the down-sampling operation and \( s \) represents the spatial scale. By experiment we found that the number of spatial scale \( S = 1 \) could achieve satisfying results.

In order to reduce illumination effects and prevent some local image regions to dominate the energy spectrum, the pre-filtering should be applied to intensity building image at different scales. Differing from the subtraction with mean intensity of the image and division with standard deviation of the intensity image [46], the proposed pre-filtering consists in a local normalization of the intensity variance

\[
p_e(x,y) = \frac{p_b(x,y) \times h(x,y)}{\varepsilon + \sqrt{p_b(x,y) \times h(x,y)}^2 \times g(x,y)}
\]

where \( g(x,y) \) is an isotropic low-pass Gaussian spatial filter with a radial cut-off frequency at 0.015 c/p (cycles/pixel) and \( h(x,y) = 1 - g(x,y) \). The numerator is a high-pass filter that cancels the mean intensity value of the image and whitens the energy spectrum at the very low spatial frequencies. The denominator acts as a local constant that avoids noise enhancement in constant image regions. We set experimentally \( \varepsilon = 20 \) for input images with intensity values in the range [0,255]. This pre-filtering stage affects only the very low spatial frequencies (below 0.015 c/p) and does not change the mean spectral signature. Compared with the local normalization method in [46], the image change after pre-filtering is shown in Fig. 1. From Fig. 1(c), we can see that the histogram is unimodal distribution after the proposed pre-filtering on the intensity image. Compared with the multimodal distribution of the histogram in Fig. 1(f), the unimodal distribution is more beneficial to later classification tasks. It is clear that the proposed pre-filtering stage achieves good normalizing result and better reduces the distribution of illumination.

After pre-filtering, for different scale intensity images, the corresponding spatial distribution of spectral information can be described by discrete Fourier transform

\[
P_s(f_x,f_y) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} p_e(x,y)h(x,y)e^{-j2\pi(f_xx+f_yy)}
\]

where \( f_x \) and \( f_y \) are the spatial frequency variables, \( h(x,y) \) is a circular Hanning window. The energy spectrum, \( A^2(f_x,f_y)^2 = |P_s(f_x,f_y)|^2 \), provides structural information that represents the spatial frequencies spread everywhere in the image and thus information about the orientation, smoothness, length and width of the contours that compose the building image. Therefore, the energy spectrum provides a building representation invariant with respect to object arrangements and object identities, encoding only the dominant structural patterns present in the image. The dimensionality of the energy spectrum is \( N \times M \).

Obviously, the energy spectrum information provides high dimensional representations of the input image. A way to reduce the dimensionality is to reduce the number of samples of the energy spectrum function. The dimensionality reduction function is defined as

\[
g_c = \langle A^2, G_i \rangle = \iint A^2(f_x,f_y)G_i(f_x,f_y)df_xdf_y
\]
G_i is a set of Gaussian functions arranged in a log-polar array and calculated by rotating and scaling the function

\[ G(f, f_r) = e^{-\frac{1}{2} \sigma^2 (e^{-\frac{1}{2} (f - f_0)^2 / \sigma^2} + e^{\frac{1}{2} f_0^2 / \sigma^2})} \]  

(7)

We use a set of Gaussians distributed in 4 frequency bands with central frequencies \( f_0 \) at 0.02, 0.04, 0.08 and 0.16 and \( c/p \) and \( c\) of orientations at each band. So, there are 32 Gabor filters of varying frequency bands and orientations. The energy spectrum \( A^2(f, f_r) \) is then represented by a series of maps \( g_s = (g_s)_i \) with \( i = 1, 2 \), \( 3 \), \( 4 \). The spectrum map \( g_s \) can be sampled with coarsely resolution, which benefit to classification tasks. This was testified in [20]. So, we decompose each \( g_s \) into \( 4 \times 4 \) grid sub-regions. Then the mean value of each sub-region is calculated for final representation, i.e. 16 mean values are utilized to represent each \( g_s \) and \( 32 \times 16 = 512 \leq (N \times M) \) values of gist feature are obtained for image representation at different scales. The whole procedure is shown in Fig. 2.

Following these procedure at different scales, we obtain a MS-gist feature of each image. Then, we reshape the MS-gist feature map in a column vector \( g \). Finally, each building image is represented by a column vector \( g \) and building dataset can be described as a MS-gist feature space.

As it is illustrated in the literatures [34,35,40–44], since the high-dimensional vector \( g \) are obtained from the building images taken from different viewpoints or under different lighting conditions, \( g \)'s of a building lie on a sub-manifold, which is extrinsically high dimensional and intrinsically low dimensional. Therefore, it is necessary for us to obtain the more compact low-dimensional representation of the gist feature manifold using more effective dimensionality reduction techniques. The following section will focus on developing a more effective linear dimensionality reduction method to enhance the performance on building recognition.

3. Enhanced fuzzy local maximal margin embedding (EFLMME)

In order to remove some redundant and useless information and alleviate computational complexity in building recognition while preserving sufficient discriminative information in the subsequent recognition stage, dimension reduction technology is applied to the MS-gist features of the building images. Here, the high-dimensional ambient space is the MS-gist feature space.

Our objective is finding an optimal mapping to project the high-dimensional MS-gist features to a low-dimensional subspace for classification.

3.1. Overview of recent works on dimensionality reduction

There are large amount of dimensionality reduction techniques developed in the literatures. Among them, there are two main fields: conventional linear dimensionality reduction techniques and graph-based manifold learning algorithms.

Two of the conventional linear dimensionality reduction methods are principal component analysis (PCA) [32] and linear discriminant analysis (LDA) [30,33], which were widely used in face image recognition and called as Eigenfaces and Fisherfaces, respectively. PCA, which is an unsupervised method, aims to preserve total variances by maximizing the trace of feature variance matrix. PCA may be unsuitable for classification tasks since the label information is not taken into account. As a supervised learning algorithm, LDA aims to maximize between-class scatter matrix and minimize the within-class scatter matrix simultaneously. Thus, LDA usually performs better than PCA in classification tasks. However, for a c-class classification task, if the dimension of the projected subspace is strictly lower than \( c-1 \), the projection to a subspace tends to merge those classes that are close together in the original feature space. Aiming to solve the class separation problem, Tao et al. [40] replaced the arithmetic mean by a general mean function and proposed a general averaged divergence analysis algorithm. By choosing different mean functions, a series of subspace selection algorithms are obtained. The geometric means for subspace selection [41], which maximizing the geometric mean of KL divergences between different class pairs are proposed. This geometric mean method achieved good performance in dealing with the class separation problem. Bian and Tao [42] introduced harmonic mean for subspace selection, which maximizes the harmonic mean of the symmetric KL divergences between all class pairs. At the same time, there are many novel graph-based manifold learning algorithms that rose in recent year. One of the most representative methods in this field should be the locality preserving projections (LPP) [34], which introduces the local geometry and neighboring relationship into the objective functions. With the similar ideas as used in LPP, Zhang et al. [43] presented a general framework, i.e. patch alignment, for dimensionality reduction and developed a new method called discriminative locality alignment (DLA) by imposing discriminative information in the part optimization stage. Zhou et al. [44] proposed the manifold elastic net (MEN), which incorporating the merits of both the manifold learning
based and the sparse learning based dimensionality reduction. Both DLA and MEN can be viewed as the graph-based dimensionality reduction methods, which construct a special graph to characterize the local structure of the dataset.

The proposed EFLMME are significantly different from the existing methods, including the recently proposed DLA and MEN. EFLMME focuses on designing two refined graphs to characterize the local compactness and local separability by introducing the fuzzy neighborhood membership degree of each sample. DLA can be viewed as a linear dimensionality reduction method which uses the local patch alignment to construct the special graph to reflect the structure of the dataset. Similar to DLA, MEN also uses the local patch alignment to characterize the manifold structure, combining the Elastic Net [47] to learn the sparse projections with interpretations.

3.2. The idea of the proposed method

In manifold learning based methods for feature extraction, samples in the margin of different classes may be projected to neighbor points in the low-dimensional space. Moreover, the graphs are constructed based on the assumption that the same membership degree are assigned to each sample in the same class or in the local neighborhood during construction of affinity graphs, which can be seen from LPP and its supervised version. Since the distances between the samples in local K-nearest neighborhood might also vary in a big range, the graphs constructed in this way might have potential disadvantages that the weights are not in accordance with the nature relations of the data in actual applications. These may lead to potential misclassification. So in this stage, we present an enhanced fuzzy local maximal marginal embedding (EFLMME) algorithm. In EFLMME, the enhanced fuzzy k-nearest neighbor is implemented in order to model the nature distribution information of original samples. Then this information is utilized to redefine the affinity weights of neighborhood graph instead of using the binary pattern or Gaussian weighted pattern. Through the fuzzy gradual pattern, two novel fuzzy gradual graphs (intra-class and interclass) are constructed. During the construction of fuzzy gradual intra-class graph, the main idea is that the nearer the neighbors, the greater the fuzzy weights. During the construction of fuzzy gradual interclass graph, the main idea is that the farther the neighbors, the greater the fuzzy weights. By doing like these, the resulting projections can map the near neighbor samples in same class becomes nearer to each other and repel the farther neighbor samples of margin between different classes to be farther from each other when they are projected to low-dimensional subspace. The details are described as follows.

3.3. Computation of enhanced fuzzy membership degree

EFLMME designs an enhanced fuzzy intra-class membership degree and interclass membership degree. For an enhanced fuzzy intra-class membership degree, k-nearest neighbors are characterized through their contributions to the intra-class compactness: the nearer the neighbors, the greater the fuzzy weights. The k-nearest neighbor in an enhanced fuzzy interclass membership degree is related to the interclass separability: the farther the neighbors, the greater the fuzzy weights.

The enhanced fuzzy intra-class membership degree is defined as below

\[
u_{ic} = \begin{cases} 
0.51 + 0.49 \times (n_{ic}/k) & \text{if the } i \text{th sample } \in \text{ cth class} \\
0.49 \times (n_{ic}/k) & \text{otherwise}
\end{cases}
\]

where

\[
n_{ic} = \sum_{q=1}^{k} (1+\varepsilon)^{k-q}; \quad 0 < \varepsilon < 0.1
\]

where \(n_{ic}\) stands for the degree of the neighbors of the ith data (pattern) belonging to the cth class and \(\varepsilon\) is a constant value that describes the compactness degree in the same classes. The value of \(\varepsilon\) is set to be 0.02. \(q\) represents the qth neighbor of the sample within the same class. The nearer the neighbor is, the more useful it is for classification. Therefore, the nearest neighbor has the biggest weight \((1+\varepsilon)^{k-1}\). The weight for the kth neighbor is 1. Based on the concept described above, it is expected that the closer samples within the same class become more compact. As a result, a more reliable intra-class membership degree description can be achieved in the low dimensional space. Thus, when the intra-class membership degree is used to construct the intra-class graph, the resulting graph can better reflect the intra-class data structure.

The enhanced fuzzy interclass membership degree is defined in a similar format of fuzzy intra-class membership degree as below

\[
p_{ic} = \begin{cases} 
0.51 + 0.49 \times (n_{ic}/k) & \text{if the } i \text{th sample } \notin \text{ cth class} \\
0.49 \times (n_{ic}/k) & \text{otherwise}
\end{cases}
\]

where

\[
n_{ic} = \sum_{q=1}^{k} (1+\beta)^{k-q}; \quad 0 < \beta < 0.1
\]

where \(\beta\) is a constant representing the separation degree in different classes. The value of \(\beta\) is set to be 0.02. To define an interclass membership degree, a further neighbor is more useful for classification. The furthest neighbor, i.e. the kth neighbor has the biggest weight \((1+\beta)^{k-1}\). Thus the far samples belonging to different classes are expected to become more separate in the low dimensional space. Thus, a more reliable interclass membership degree description is obtained and can be used to create an interclass graph that can better reflect the interclass data structure.

3.4. Construction of fuzzy gradual neighborhood graphs

Suppose that the mapping from \(x_i\) to \(y_i\) is \(\omega = [\omega_1, \omega_2, ..., \omega_d]\), i.e. \(y_i = \omega^T x_i\).

By following the graph embedding formulation, interclass separability, namely margin, is characterized by a fuzzy gradual interclass graph

\[
S_k = \sum_{y_i} \| \omega^T x_i - \omega^T x_{k}^2 \| \| W_{ij} \| = 2 \left( \sum_{i} \omega^T x_i D_i^k \omega - \sum_{i} \omega^T x_i W_{i,j} \omega \right)
\]

\[
= 2(\omega^T X D^k \omega - \omega^T X W \omega) = 2(\omega^T X (D^k - W^k) X^T \omega),
\]

\[
W_{ij} = \begin{cases} 
p_{i,j} & \text{if } i \in \mathcal{N}_k(j), j \in \mathcal{N}_k(i), j \in \text{ Class } \mathcal{m} \text{ or } i \in \mathcal{N}_k(j), i \in \text{ Class } \mathcal{m} \\
0, & \text{else}
\end{cases}
\]

where \(W^k\) characterizes the affinity weights between classes, in which each elements \(W_{ij}\) refers to the weight of the edge between \(x_i\) and \(x_j\) in different classes, \(D^k\) is a diagonal matrix with diagonal elements \(D_{ii}^k = \sum W_{ii}\) and \(\mathcal{N}_k(i)\) indicates the index set of the k nearest neighbors of the sample \(x_i\) in different classes.
Intra-class compactness is calculated from enhanced fuzzy intra-class graph

\[
S_{C} = \sum_{\eta} |\omega^{T}x_{\eta} - \omega^{T}x_{\eta}|^{2}W_{\eta}^{\rho} = 2 \left( \sum_{\eta} \omega^{T}D_{\eta}^{\rho}x_{\eta}^{T}\omega - \sum_{\eta} \omega^{T}x_{\eta}^{T}W_{\eta}^{\rho}x_{\eta}^{T}\omega \right)
\]

\[
= 2(\omega^{T}DD^{T}\omega - \omega^{T}DW^{C}\omega) = 2\omega^{T}(D^{T} - W^{C})\omega
\]

\[(12)\]

\[
W_{ij}^{\rho} = \begin{cases} \mu_{i,m} \times \mu_{j,m}, & \text{if } i \in N_{k_{2}}(j) \text{ or } j \in N_{k_{2}}(i) \text{ and } i,j \in \text{Class } m \\ 0, & \text{else} \end{cases}
\]

where \(W^{C}\) characterizes the affinity weights in the same class, in which each element \(W_{ij}^{\rho}\) refers to the weight of the edge between \(x_{i}\) and \(x_{j}\) in the same class, \(D^{T}\) is a diagonal matrix with diagonal elements \(D_{\eta} = \sum_{\eta} W_{\eta}^{\rho}\) and \(N_{k_{2}}(i)\) indicates the index set of the \(k_{2}\) nearest neighbors of the sample \(x_{i}\) in the same class.

3.5. Objective function

The proposed algorithm is expected to find the optimal projections that can minimize enhanced fuzzy intra-class graph and simultaneously maximize enhanced fuzzy interclass graph. We then have following constrained optimized problem:

Maximize \(S_{C} = \sum_{\eta} \sum_{j} |\omega^{T}x_{\eta} - \omega^{T}x_{j}|^{2}W_{\eta}^{\rho}\)

subject to \(S_{C} = \sum_{\eta} \sum_{j} |\omega^{T}x_{\eta} - \omega^{T}x_{j}|^{2}W_{\eta}^{\rho} = 1\)

\[(14)\]

The above criterion is formally similar to the Fisher criterion since they are both Rayleigh quotient problems. Therefore, we can obtain its optimal solutions by solving a generalized eigen-equation

\[
X(D^{T} - W^{C})X^{T}\omega = \lambda X(D^{T} - W^{C})X^{T}\omega
\]

where \(\omega\) is generalized eigenvector corresponding to generalized eigenvalue \(\lambda\). Then, we can select the eigenvectors associated to the \(d\) largest eigenvalues as the optimal projection matrix \(P_{\text{EFLMME}}\), where \(P_{\text{EFLMME}} = [\omega_{1}, \omega_{2}, ..., \omega_{d}]\).

3.6. Procedure of MS-gist feature manifold model

The proposed algorithm is described as a two-stage model. In the first stage, MS-scale gist feature is extracted. In the second stage, EFLMME is applied to multi-scale gist feature space for dimensionality reduction. The following steps summarize the proposed algorithm:

Step 1. Convert the input color image to intensity image using Eq. (1) and decompose the intensity image at different scales using Eq. (3).

Step 2. Filter the intensity image at different scales using Eq. (4).

Step 3. Compute the MS-gist feature using Eqs. (5) and (6).

Step 4. Compute the fuzzy gradual intra-class and the inter-class membership degree matrices on the MS-gist feature space using Eqs. (8) and (9), respectively.

Step 5. Compute affinity weights using Eqs. (11) and (13), respectively. An edge is added between \(x_{i}\) and \(x_{j}\) from the same class if \(x_{i}\) is one of \(x_{j}\)'s \(k\)-nearest neighbors. For the interclass graph, \(x_{i}\) is connected with \(x_{j}\) from the different classes if \(x_{i}\) is one of \(x_{j}\)'s \(k\)-nearest neighbors.

Step 6. Create the fuzzy gradual intra-class compactness graph and fuzzy gradual interclass separability graph using Eqs. (10) and (12), respectively.

Step 7. Solve the generalized eigen-function of Eq. (15) and obtain the optimal projection matrix \(P_{\text{EFLMME}}\).

Once the project matrix \(P_{\text{EFLMME}}\) is obtained through using the FLMME algorithm, the nearest neighbor classification can be used for classification.

4. Experiments and discussions

To evaluate performance of our proposed model, experiments were carried out on Sheffield buildings database [48], compared with other models: (a) the visual gist based building recognition model (VGBR): each image is described by a gist feature based on visual features [17]; (b) the hierarchical building recognition model (HBR): each image is represented by a local color histograms indexing vector [16]. At the same time, EFLMME is evaluated on Sheffield buildings database compared with PCA, LDA, LPP and DLA.

Sheffield buildings database includes 40 buildings, i.e. 40 categories in total. For each building, the number of images varies from 35 to 334. In summary, there are 3192 images of size 120 × 160. The building database is of a variety of challenges, i.e. rotation, variant lighting conditions, viewpoint changes, occlusions and vibration.

They were taken at different times on separate days. Different times cover early morning, noon, mid-afternoon and early evening. This leads to highly variable lighting conditions and results in the building recognition task more challenging. Buildings include churches and a variety of modern buildings, such as exhibition halls. Images for each building were taken from different viewpoints, varying from three to nine views; through moving the camera from side to side, video clips are also obtained with multiple views. Buildings might be occluded from trees, moving vehicles, other buildings or themselves. Furthermore, some videos were captured by walking from one side to another, which creates the additional challenge of movement/camera-shake.

For each building, the number of the images is significantly different in the original database. In order to test the performance of the proposed framework with the small and the large number of training samples, two subsets (i.e. subsets I and II) were selected from the Sheffield building database and used in our experiment, respectively. The subset I consists 40 building categories, and the first 18 images for each building in the original database were selected and used in the experiments. Thus there were 720 images in the subset I in total. Some sample images in the subset I are shown in Fig. 3. The subset II consists 39 building categories and the first 60 images for each building were selected and used in the following experiments. The total number of images in subset II is 2340. Some sample images in the subset II are shown in Fig. 4.

In the proposed model, we first converted the input RGB color image to intensity image using Eq. (1) and decomposed the intensity image at different scales using Eq. (3). For each intensity image with different scales, we filtered the intensity image using Eq. (4) and then extracted the MS-gist feature of each building image using Eqs. (5) and (6). But the MS-gist feature space still contains redundant and useless information and the dimensions of the feature are still high. In order to remove this redundant information and alleviate the computational complexity in classification, we then use the proposed EFLMME algorithm for further dimensionality reduction. In this stage, we computed the fuzzy gradual intra-class and the interclass membership degree matrices on the MS-gist feature space using Eqs. (8) and (9), and then created the fuzzy gradual intra-class compactness
graph and fuzzy gradual interclass separability graph using Eqs. (10) and (12). Subsequently, we solved the generalized eigen-function of Eq. (15) and obtained the optimal projection matrix. Finally, the nearest neighborhood classifier was utilized for classification. The number of intra-class local neighborhood was chosen as \( k_1 = l - 1 \) in the EFLMME, where \( l \) denotes the number of training samples per category. The choice is testified to be reasonable in the observation space in [45].

4.1. Experiments on the subset I

The subset consists 40 building categories. There are 18 images for each building and total number is 720. In the experiments, each image of the subset I is firstly represented by a 512 \( \times \) 2 MS-gist feature, then the feature is reshaped as a 1024-dimensionality vector, and feature dimensionality is subsequently reduced by PCA, LDA, LPP, DLA and the proposed algorithm (EFLMME).

4.1.1. Validation of the proposed MS-gist feature

In this subsection, we use the first \( l \) (\( l = 3, 5, 7, 9 \)) images of each building for training and the rest for testing. In order to validate the performance of the MS-gist feature, the nearest neighbor (NN) classifier is directly applied to the MS-gist feature space and compared with raw data space, HBR feature space and VGBR feature space. The results are showed in Table 1. The performance of the proposed MS-gist feature is superior to those of other model. Since pre-filtering steps and the holistic structure information of building images at different scales are introduced in the MS-gist model, the extracted MS-gist features are more robust to different viewpoints, different lighting conditions and occlusions. Thus, the proposed model obtains better performance for building recognition.

<table>
<thead>
<tr>
<th>Model</th>
<th>3 train</th>
<th>5 train</th>
<th>7 train</th>
<th>9 train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data space + NN</td>
<td>45.67 (98)</td>
<td>61.15 (105)</td>
<td>67.27 (95)</td>
<td>71.94 (103)</td>
</tr>
<tr>
<td>HBR feature space + NN</td>
<td>63.26 (96)</td>
<td>78.24 (103)</td>
<td>83.26 (96)</td>
<td>87.56 (101)</td>
</tr>
<tr>
<td>VGBR feature space + NN</td>
<td>64.17 (100)</td>
<td>80.81 (90)</td>
<td>86.72 (98)</td>
<td>89.97 (103)</td>
</tr>
<tr>
<td>MS-gist feature space + NN</td>
<td>75.33 (84)</td>
<td>88.36 (90)</td>
<td>89.18 (98)</td>
<td>93.33 (105)</td>
</tr>
</tbody>
</table>

4.1.2. Validation of the different feature plus EFLMME

In this subsection, we use the first \( l \) (\( l = 3, 5, 7, 9 \)) images of each building for training and the rest for testing. To demonstrate the
benefits of the proposed feature extraction model, we compared it with raw building data and other models [16,17]. In the second stage, the identical algorithm, i.e. EFLMME, is used to the features of different models for dimensionality reduction. Table 2 lists the best performances obtained by different feature model. As can be seen in Table 2, the proposed model performs better than the other models when the same dimensionality reduction method is used. Furthermore, when compared Table 1 with Table 2, we can find EFLMME help all feature models improve the building recognition rates. Therefore, not only the MS-gist feature can more effective for building recognition, but also the EFLMME can further improve the classification accuracies.

4.1.4. Random experiment

In this subsection, for each building, I (=3,5,7,9) images are randomly selected as training samples, and the rest images are used for test. All the five methods, PCA, LDA, LPP, DLA and EFLMME are utilized for dimensionality reduction on the multi-scale gist feature space. The procedure is repeated 50 times and the average recognition rate was calculated. In general, the recognition rates vary with the dimension of the multi-scale gist feature space. Fig. 5 shows the variation of average recognition rates using different algorithms with different dimensions. It can be seen from the figures that the performance of EFLMME outperforms the other methods on the MS-gist feature space.

4.2. Experiments on the subset II

To further evaluate the performance of the proposed model, experiments were carried on the bigger dataset, i.e. the subset II. The subset II consists 39 building categories. There are 60 images for each building and the total number is 2340. Similarly, each image is firstly represented by a S12 x 2 MS-gist feature, and the dataset is represented by the MS-gist features. Subsequently, techniques of dimensionality reduction are applied to the multi-scale gist feature manifold.

4.2.1. Validation of the proposed MS-gist feature

In this subsection, for each building, we use the first I (=10,20,30,40) images for training and the rest for test. In order to validate the performance of the MS-gist features, the nearest neighborhood classifier is directly applied to the MS-gist feature space and compared with raw data space, HBR feature space and VGBR feature space. The results are showed in Table 4. The performance of the proposed MS-gist feature is superior to those of other feature.

4.2.2. Validation of the different feature plus EFLMME

In this subsection, for each building, we use the first I (=10,20,30,40) images for training and the rest for testing. To demonstrate the benefits of the proposed feature extraction model, we compared it with raw building data and other models in [16,17]. The identical algorithm--EFLMME is used for dimensionality reduction. In general, the recognition rates varied with the dimensions of the features. The best performances obtained by different models are presented in Table 5. Fig. 6 demonstrates the recognition accuracy of raw data model, HBR model, VGBR model and the proposed model over the variations of the dimensionality corresponding to the first I (=10,20,30,40) images for training. From Table 5 and Fig. 6, we can see that the performance of the proposed MS-gist features also outperforms other features significantly. This testifies the effectiveness of the proposed model again.

### Table 2

<table>
<thead>
<tr>
<th>Feature Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data model + EFLMME</td>
<td>51.74 (31) 68.67 (33) 75.32 (49) 80.48 (35)</td>
</tr>
<tr>
<td>HBR model + EFLMME</td>
<td>65.15 (81) 79.37 (77) 84.49 (97) 89.89 (95)</td>
</tr>
<tr>
<td>VGBR model + EFLMME</td>
<td>66.50 (87) 82.88 (101) 87.50 (101) 91.11 (99)</td>
</tr>
<tr>
<td>MS-gist model + EFLMME</td>
<td>80.50 (81) 91.73 (103) 93.64 (99) 97.61 (79)</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Feature Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>75.33 (84) 88.36 (90) 89.18 (98) 93.33 (105)</td>
</tr>
<tr>
<td>PCA</td>
<td>76.46 (93) 89.15 (105) 91.18 (105) 94.61 (95)</td>
</tr>
<tr>
<td>LDA</td>
<td>74.83 (29) 89.04 (39) 90.23 (39) 92.22 (37)</td>
</tr>
<tr>
<td>LPP</td>
<td>70.50 (53) 82.88 (51) 87.72 (53) 88.89 (53)</td>
</tr>
<tr>
<td>DLA</td>
<td>78.67 (52) 91.27 (56) 92.23 (86) 95.78 (74)</td>
</tr>
<tr>
<td>EFLMME</td>
<td>80.50 (81) 91.73 (103) 93.64 (99) 97.61 (79)</td>
</tr>
</tbody>
</table>

MS-gist features. Compared to EFLMME for dimension reduction in the second stage, the extracted MS-gist features in the first stage makes greater contributions to total recognition accuracies.
results are showed in Table 6. As can be seen in Table 6, the PCA, LDA, LPP and DLA. For each building, we also use the first and the raw data space in the second stage and compared with

4.2.3. Global validation of the proposed model

Recognition accuracy (%) on the subset II and the corresponding dimension Table 5

Table 5

<table>
<thead>
<tr>
<th></th>
<th>10 train</th>
<th>20 train</th>
<th>30 train</th>
<th>40 train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data space + NN</td>
<td>54.72 (92)</td>
<td>57.62 (95)</td>
<td>60.43 (99)</td>
<td>63.59 (105)</td>
</tr>
<tr>
<td>HBR feature space + NN</td>
<td>71.43 (94)</td>
<td>73.33 (88)</td>
<td>76.23 (91)</td>
<td>79.19 (103)</td>
</tr>
<tr>
<td>VGBR feature space + NN</td>
<td>73.38 (90)</td>
<td>77.81 (88)</td>
<td>81.82 (103)</td>
<td>82.77 (100)</td>
</tr>
<tr>
<td>MS-gist feature space + NN</td>
<td>79.05 (85)</td>
<td>87.14 (92)</td>
<td>91.62 (105)</td>
<td>93.85 (95)</td>
</tr>
</tbody>
</table>

4.3. Discussions

Based on the experimental results in Sections 4.1–4.2, we can obtain the following conclusions:

(1) As far as building recognition tasks is concerned, the proposed model achieves remarkable improvements when compared with other models. This shows that the proposed model is an effective method in practice of building recognition. Experimental results show that the proposed model can handle the building recognition problem caused by rotation, variant lighting conditions, and occlusions very well.

(2) In this paper, the proposed model is described in two stages: extraction of MS-gist feature and dimensionality reduction using EFLMME. From Tables 1 and 4 and Fig. 6, we can see that performances on the MS-gist feature space are remarkably better than other feature space. From Tables 2 and 5, we can see that EFLMME using on the different feature spaces can greatly enhance the recognition rates. This indicates EFLMME is an effective dimensionality reduction method. As is shown in Tables 3 and 6, the MS-gist feature achieves a better result compare with the raw data feature under the different dimensional reduction methods.

(3) From Tables 3 and 6 and Figs. 5 and 7, we also find the performance of EFLMME is superior to PCA, LDA, LPP and DLA both on the MS-gist feature space and raw data space. The reason is that the enhanced fuzzy membership degree can efficiently handle the vagueness and ambiguity of samples degraded by poor illumination, shape and view-point variations. In other words, the enhanced fuzzy membership degree helps pull the near neighbor samples in same class farther and repel the far neighbor samples of margin between different classes farther and farther. So, the fuzzy gradual graphs based on the fuzzy gradual membership degree can better characterize the compactness and separability. Thus, EFLMME performs better than other dimensionality reduction techniques.

(4) Our experimental results show that the proposed model has great superiority to other models. The reasons may be that the proposed two stage method can provide more separate intrinsic feature of the different buildings. In the first stage, the MS-gist feature provides structural information represents the spatial frequencies spread everywhere in the image and thus information about the orientation, smoothness, length and width of the contours that compose the building images. Therefore, the MS-gist feature effectively captures the building representation,
which is invariant with rotations, variant lighting conditions, and occlusions. In the second stage, the performance of EFLMME outperforms those of PCA, LDA, LPP and DLA. This indicates the EFLMME is an effective dimensionality reduction method, which performs dimensionality reduction on the proposed MS-gist feature and preserves more discriminative information for building recognition. Therefore, after EFLMME applied to the MS-gist feature space, the recognition rate of the proposed model can be further increased.

5. Conclusions

In this paper, we have described a MS-gist feature for building recognition and proposed a new approach called EFLMME for further improve the building recognition rate. The MS-gist features can be stable to capture the representation features of the building images with rotation, variant lighting conditions and occlusions. In EFLMME the enhanced fuzzy membership degree helps pull the near neighbor samples in same class nearer and nearer and repel the far neighbor samples of margin between different classes farther and farther. Therefore, the fuzzy gradual graphs, which are based on the fuzzy gradual membership degree, can better characterize the compactness and separability. Experiments were carried out on Sheffield buildings database compared with the VGBR model and HBR model. At the same time, EFLMME is evaluated on Sheffield buildings database and compared with PCA, LDA, LPP and DLA. The results show that the proposed model is an effective method in building recognition and can handle the

---

### Table 6

<table>
<thead>
<tr>
<th>Multi-scale gist feature space</th>
<th>Raw data space</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 train</td>
<td>20 train</td>
</tr>
<tr>
<td>NN</td>
<td>79.05 (85)</td>
</tr>
<tr>
<td>PCA</td>
<td>79.15 (71)</td>
</tr>
<tr>
<td>LDA</td>
<td>79.72 (35)</td>
</tr>
<tr>
<td>LPP</td>
<td>73.72 (19)</td>
</tr>
<tr>
<td>DLA</td>
<td>82.00 (50)</td>
</tr>
<tr>
<td>EFLMME</td>
<td>84.85 (43)</td>
</tr>
</tbody>
</table>

---

Fig. 6. (a)–(d) are the recognition accuracy with the dimension of raw data model, HBR model, VGBR model and the proposed model on subset II corresponding to the first \( l \) (\( l=10,20,30,40 \)) images for training. (a) The first 10 images for training; (b) the first 20 images for training; (c) the first 30 images for training; (d) the first 40 images for training.
building recognition problem caused by rotations, variant lighting conditions and occlusions very well.

However, we mainly focus on the building recognition accuracy rate in the paper instead of the truly semantic concept extraction in this paper. So the MS-gist feature may neglect to discover a truly semantic concept of the image, which is probably much important in the long run. Moreover, EFLMME is a supervised learning on building recognition, the label information is needed to find, which may be a boring work. Therefore, in the future, we will focus both on the truly semantic feature extraction and the unsupervised learning algorithm with good performance on building recognition.

Acknowledgments

This work is partially supported by the Fujian Provincial Department of Science and Technology of China under Grant nos. JK2010046 and JB10135. It is also partially supported by the National Science Foundation of China under Grant nos. 60472061, 60632050 and 90820004 and Hi-Tech Research and Development Program of China under Grant no. 2006AA042238.

References


