Building Recognition Using Local Oriented Features

Jing Li and Nigel Allinson, *Senior Member, IEEE*

***Abstract*—Building recognition is an important task for a wide range of computer vision applications, e.g., surveillance and intelligent navigation aid. However, it is also challenging since each building can be viewed from different angles or under different lighting conditions, for example, resulting in a large variability among building images. A number of building recognition systems have been proposed in recent years. However, most of them are based on a complex feature extraction process. In this paper, we present a new building recognition model based on local oriented features with an arbitrary orientation. Although the newly proposed model is very simple, it offers a modular, computationally efficient, and effective alternative to other building recognition techniques. According to a comparison of experimental results with the state-of-the-art building recognition systems, it is shown that the newly proposed SFBR model can obtain very satisfactory recognition accuracy despite its simplicity.**

***Index Terms*—Building recognition, dimensionality reduction, local oriented features, max pooling, steerable filters.**

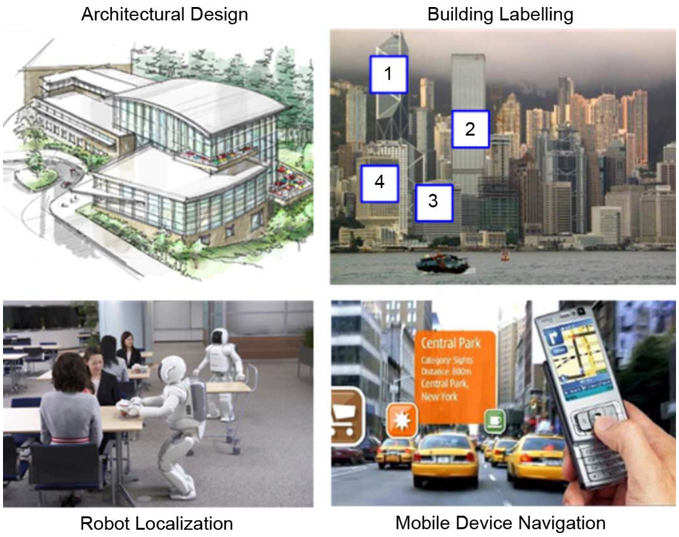


Fig. 1. Different applications of building recognition.

I. Introduction

BUILDING recognition, an intra-class recognition task, is aiming at distinguishing different buildings in a large-scale image database. It becomes an important yet challenging task and has attracted considerable attention in computer vision research. Building recognition can be used in various kinds of applications (as shown in Fig. 1), including surveillance [7], automatic target detection and tracking [47], architectural design, building labeling in videos, 3-D city reconstruction, real-time mobile device navigation [1], [20], [37], robot motion representation [29], and robot localization [45]. Take robot localization as an example. For a mobile robot moving between buildings, robot localization is a process to determine its accurate position from data collected from intelligent sensors, e.g., inertial [50], RF, laser, and ultrasonic range sensors [25]. Each type of sensors works well for a certain condition. For example, ultrasonic range sensors are useful in operating within a mapped environment containing known obstacles while inertial sensors cannot satisfactorily deal with drift, which is accentuated when the robot moves continuously over a long time. Nevertheless, the results of multiple sensors can be combined for robot operations in a wide range of conditions. After data acquisition, vision-based techniques should be able to reliably recognize different building images from various viewpoints. From this perspective, building recognition plays an important role in real-world localization and navigation applications.

Compared to general object recognition, the task of building recognition is more challenging since images for each building contain different amount of variability—they may be taken from different viewpoints, under different lighting conditions, or suffer from partial occlusions from trees, moving vehicles, other buildings, or themselves. How to deal with these challenges is an interesting research problem and a small number of building recognition systems [1], [12], [15], [20], [21], [27], [28], [42] have been proposed in the last few years.

Generally, most of the existing building recognition systems adopt a complex feature extraction process to represent an image. For example, both global features (e.g., color [16], [31], texture [31], and shape [16]) and local features (e.g., SIFT [30] and shape contexts [4]) are integrated to obtain satisfactory performance. Using more features may bring better results [11], however, it also means the feature representation requires more computational cost and is not easy to implement. Aware of this, we investigated whether there is a simple way for feature extraction in the building recognition task. In this paper, we propose a simple building recognition model, called steerable filtered-based building recognition (SFBR) model which only uses a type of local oriented features in representing an image. Although steerable filters [14] have been previously utilized for other recognition tasks, it is the first time they have been

Manuscript received September 29, 2012; revised January 16, 2013; accepted January 25, 2013. Date of publication February 08, 2013; date of current version August 16, 2013. This work was supported in part by the National 863 Program under Grant 2013AA013804, the National Natural Science Foundation of China under Grant 61175072 and Grant 51165033, the National Basic Research Program of China under Grant 2011CB302400, and the Science and Technology Department of Jiangxi Province of China under Grant 20121BBE50023. Paper no. TII-12-0694.

J. Li is with the School of Information Engineering and Jiangxi Provincial Key Laboratory of Intelligent Information Systems, Nanchang University, Nanchang 330031, China (e-mail: jing.li.2003@gmail.com).

N. Allinson is with the School of Computer Science, University of Lincoln, Lincoln LN6 7TS, U.K. (e-mail: nallinson@lincoln.ac.uk).

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

Digital Object Identifier 10.1109/TII.2013.2245910

applied to building recognition. The SFBR model consists of four parts, which are: i) feature representation; ii) feature pooling; iii) dimensionality reduction; and iv) classification. In the following paragraphs, we give the motivations and describe each part of this model in more detail.

Feature representation usually serves as the first step in describing different objects in an image. In the main, it contains two kinds of representation: global feature representation and local feature representation. Global feature representation regards an image as a whole and is able to express its global appearance by extracting features at each pixel of an image. Local features [4], [26], [30], [34], referring to image patterns that are different from those in their neighborhoods, could describe local information of an image and thus are invariant to small image transformations. As we mentioned before, the building recognition task entails a few challenges, e.g., scaling and occlusions. Considering this, types of features that are distinctive while robust to different geometric and photometric transformations are a necessity—that is exactly the promising properties and capabilities of local features [26]. Moreover, due to the characteristics of the components of a building (e.g., windows, doors, and bricks), local features that can deal with edge information with varying angles offer the most potential. To this end, we chose steerable filters [14] from a number of local features because they are able to select oriented features, which is observable in mammalian cortical cells and similar to the visual processing in our receptive fields. Steerable filters are a set of basis filters that can synthesize any filter with an arbitrary orientation, where second-order, third-order, and fourth-order are widely utilized.

To achieve invariance to small shifts in position and changes in lighting conditions, we use feature pooling [5], [41] to preserve discriminative information while discarding irrelevance by combining feature responses over nearby regions into a more useful statistical representation. The combination rule can be a sum, an average, or a max. Max pooling is a key mechanism for object recognition in the cortex. Therefore, in SFBR, it is used to search the max value of the steerable responses over local patches since its efficiency has been demonstrated in [38].

Nevertheless, the extracted features are of high dimensions, which may contain some redundant information. Directly utilizing these high-dimensional features for subsequent operations, i.e., classification, not only results in high computational cost and requires large memory, but also encounters the curse of dimensionality [3]. Fortunately, dimensionality reduction (DR) can make data more compact for representation and alleviate these problems. Among a large number of DR techniques, linear subspace methods [22], [33] and manifold learning algorithms [2], [9], [19], [22], [33], [39], [48] dominate. Linear subspace methods project the original higher dimensional data points into a lower dimensional space by a linear transformation; while manifold learning algorithms aim to explore the local geometrical structure in the low-dimensional manifold embedded in the high-dimensional space. In SFBR, we apply linear discriminant analysis (LDA) [33], one of the most important linear subspace methods, for dimensionality reduction since LDA is a discriminative model which separates data points into different classes in the projected lower subspace and thus suitable for the building recognition task. Furthermore, to evaluate the effectiveness of different dimensionality reduction methods, we utilize two of representative manifold learning algorithms to substitute LDA, which are locality preserving projections (LPPs) [19] and discriminative locality alignment (DLA) [48].

After dimensionality reduction, we use a support vector machine (SVM) [16], [17] to discriminate different buildings. An SVM is a binary classifier which maximizes the margin between two classes. Because of its good generalization abilities and no requirement for prior knowledge about the data, it has been universally utilized as one of the most popular learning machines in various research areas, e.g., face recognition [17] or texture classification [23]. Here, we adopt a linear SVM for simplicity.

We evaluate the performance of SFBR on the Sheffield building image dataset (SBID) [27] which contains various kinds of challenges. Furthermore, we compare its performance with state-of-the-art building recognition systems, namely the hierarchical building recognition (HBR) system [49] and the biological-plausible building recognition (BPBR) scheme [27]. Through experiments, we offer evidence that although SFBR is the simplest, it performs best and produces a very good recognition rate.

This paper contributes to the efforts of developing simple, practical, modular, and easy-to-implement building recognition algorithms that are both cost and computationally efficient. In addition to effective local features, the proposed SFBR takes advantage of dimensionality reduction and feature pooling to recognize different buildings in a complicated image database. The remainder of this paper is organized as follows. In Section II, we review related work on building recognition. In Section III, we present the newly proposed SFBR model in detail. In Section IV, we evaluate the performance of SFBR and give the comparison results among the state-of-the-art building recognition systems. Section V concludes the paper and provides some discussion.

II. Related Work

Existing building recognition systems can be roughly divided into three categories: 1) clustering-based methods; 2) feature representation-based algorithms; and 3) others. Clustering-based methods aim to discover the relationships among different image structures by grouping them into different clusters; feature representation-based algorithms focus on the process of feature extraction in building recognition. In the following sections, representative building recognition systems will be briefly reviewed.

A. Clustering-Based Methods

To explore semantic relationships among different low-level visual features, principles of the perceptual grouping [21] were utilized to hierarchically extract various image structures, e.g., straight line segments for content-based image retrieval (CBIR) of buildings.

In [28], color, orientation, and spatial information for each line segment were integrated and grouped into a type of mid-level feature, i.e., consistent line clusters, where the intracluster and intercluster relationships were utilized to recognize different buildings.

Zhang and Kosecká [49] proposed a building recognition system based on vanishing point detection and localized color histograms. Detected line segments are grouped into dominant vanishing directions and vanishing points are estimated by the expectation maximization (EM) algorithm. After that, image pixels satisfying some certain constraints will be divided into three groups, namely left, right, and vertical, and localized color histograms will only be computed on these pixels. Because of the fast indexing step using localized color histograms, this method achieved some improvement in efficiency and has attracted the most attention.

Chung *et al.* [12] utilized sketch-based representations to find repetitive components of a building, e.g., windows and doors, for office-building recognition. The scheme detects multiscale maximal stable extremal regions (MSERs) [32] and describes the normalized MSER patches using histogram of oriented gradients. Afterwards, k-means clustering is applied to grouping the local patches into different structural components and spectral graph matching is conducted to find corresponding clusters between a query image and a reference image in the database. However, this method only focuses on office-building recognition while its performance for other building types has not been demonstrated.

B. Feature Representation-Based Algorithms

Hutchings and Mayol [20] designed a building recognition system for mobile devices to serve as a tourist guide in the world space. Given a query image, its local features are extracted and described by the Harris corner detector [18] and the SIFT descriptor [30], respectively. In the matching process, a scale is selected for each query image according to its GPS position. This results in the reduction of search space and the computational cost. However, the system fails in dealing with very large viewpoint changes.

Groeneweg *et al.* [15] implemented fast offline building recognition based on intensity-based region detection [44] and PCA [22]. Rather than fitting an ellipse around each detected region, each region is resized and represented by its RGB color values. After normalization by the sum of intensities of all pixels in the region, PCA is applied and the features are grouped into clusters, each of which is characterized by its centroid. Then, a weighted majority voting is adopted for distance measure. This approach reduces the computational cost and the storage capacity in a mobile phone platform, but it is sensitive to illumination colors and is not invariant to rotation.

In [1], a rapid window detection and localization method for buildings was put forward for mobile vision systems, where window detection, integrating line grouping, pattern detection, and gradient setting, is considered as a pattern recognition task. It is based on the extraction of multiscale Harr-like features followed by a learning stage using a classifier cascade through Adaboost [6]. The advantage of the proposed method is: instead of detecting every window of a building, only a fraction of discriminative windows need to be detected, enabling fast indexing of buildings from mobile imagery in the urban environment.

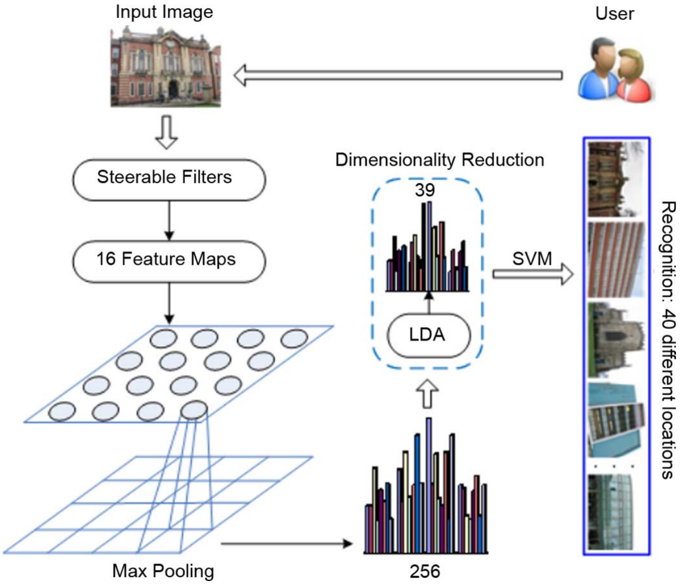


Fig. 2. Newly proposed building recognition scheme.

C. Others

Most of building recognition systems do not consider the process of human vision perception and also suffer from the curse of dimensionality due to the high dimensions of extracted features. To address these problems, Li and Allinson [27] proposed a biologically-plausible building recognition scheme which integrates biologically inspired feature extraction [36], [40] and dimensionality reduction [33].

While the above-mentioned building recognition algorithms focus on single-building recognition, i.e., each image only contains one dominant building, Trinh *et al.* [42] proposed a method to recognize multiple buildings in an image for robot intelligence. Facets of each building are extracted based on line segments and vanishing points detection, and then SIFT features [30] are utilized to describe each building.

III. Steerable Filter-Based Building Recognition (SFBR) Model

Motivated by the effectiveness of local features, we propose a SFBR scheme which is able to deal with both geometric and photometric transformations. It consists of the following stages: 1) feature representation; 2) feature pooling; 3) dimensionality reduction; and 4) classification. The newly proposed SFBR model is illustrated in Fig. 2 and each component will be introduced in the following sections.

A. Feature Representation

In SFBR, feature representation is based on the calculation of steerable filter responses at various orientations. Given an image as a function $I(\vec{p})$, where the domain of an image I is a set of pixel locations $\vec{p} = [x, y]^T$, steerable filters [26] will be defined as follows. Steerable filters [14] are a set of basis filters which can synthesize any filter F^θ with an arbitrary orientation θ, i.e., $F^\theta = \sum_{i=1}^{N} k_i(\theta) F_i$, where F_i is the ith basis filter and $k_i(\theta)$ is the linear combination coefficient. A quadrature pair of filters, which means two filters have identical frequency but one is the

Hilbert transform of the other, can be applied to synthesizing filters of a given frequency response with arbitrary phase. The derivatives of Gaussian have been demonstrated to be effective in many early vision and image processing tasks. Here, we utilize the second-order Gaussian G_2 and its corresponding Hilbert transform H_2 as the quadrature pair of filters with an identical frequency with the original parameter-setting in [14]

$$G_2(x,y) = 0.9213\left(2x^2-1\right)\exp\left(-\left(x^2+y^2\right)\right) \quad (1)$$

$$H_2(x,y) = \left(-2.205x+0.9780x^3\right) \times \exp\left(-\left(x^2+y^2\right)\right). \quad (2)$$

According to [14], the basis filters for G_2 are

$$G_{2a} = 0.9213\left(2x^2-1\right)\exp\left(-\left(x^2+y^2\right)\right) \quad (3)$$

$$G_{2b} = 1.843xy\exp\left(-\left(x^2+y^2\right)\right) \quad (4)$$

$$G_{2c} = 0.9213\left(2y^2-1\right)\exp\left(-\left(x^2+y^2\right)\right) \quad (5)$$

and the corresponding linear combination coefficients are

$$k_a(\theta) = \cos^2(\theta) \quad (6)$$

$$k_b(\theta) = -2\cos(\theta)\sin(\theta) \quad (7)$$

$$k_c(\theta) = \sin^2(\theta). \quad (8)$$

For H_2, the basis filters are

$$H_{2a} = 0.9780\left(-2.254x+x^3\right)\exp\left(-\left(x^2+y^2\right)\right) \quad (9)$$

$$H_{2b} = 0.9780\left(-0.7515+x^2\right)y\exp\left(-\left(x^2+y^2\right)\right) \quad (10)$$

$$H_{2c} = 0.9780\left(-0.7515+y^2\right)x\exp\left(-\left(x^2+y^2\right)\right) \quad (11)$$

$$H_{2d} = 0.9780\left(-2.254y+y^3\right)\exp\left(-\left(x^2+y^2\right)\right) \quad (12)$$

and the corresponding linear combination coefficients are

$$k_a(\theta) = +\cos^3(\theta) \quad (13)$$

$$k_b(\theta) = -3\cos^2(\theta)\sin(\theta) \quad (14)$$

$$k_c(\theta) = +3\cos(\theta)\sin^2(\theta) \quad (15)$$

$$k_d(\theta) = -\sin^3(\theta). \quad (16)$$

In SFBR, the second-order filter responses are computed at eight different orientations, i.e., $\pi d/4$ with $d = 0,1,2,3,4,5,6,7$, resulting in 16 feature maps for each image. Fig. 3 gives an example: the top two rows represent second-order Gaussian responses and the bottom two rows correspond to their Hilbert transform responses.

The reasons we select second-order steerable filters instead of fourth-order are as follows. Compared with a second-order filter, a fourth-order filter enables a narrower orientation tuning and thus can provide a higher-resolution analysis of orientation to discriminate multiple orientations at each location in the input patch. Nevertheless, they may give incorrect responses and generate interference effects in the energy form. Moreover, steering an N-th order polynomial requires $N+1$ basis functions [14]. In the case of a second-order steerable filter, steering G_2 needs three basis functions and steering H_2 (using a third polynomial for approximation to fit the least-squares to a polynomial times a Gaussian) needs four basis functions, that is, seven basis functions of G_2 and H_2 are sufficient to shift G_2 to arbitrary orientations. For a fourth-order filter, steering G_4 to any orientation requires 11 basis functions (five basis functions for G_4 and six basis functions for H_4). In practice, the most useful filters are those that possess a small number of basis functions, and thus higher order steerable filters are not always applicable. Furthermore, the aim of this paper is to recognize buildings in an efficient and simple way, using fourth-order filters will result in additional computational cost.

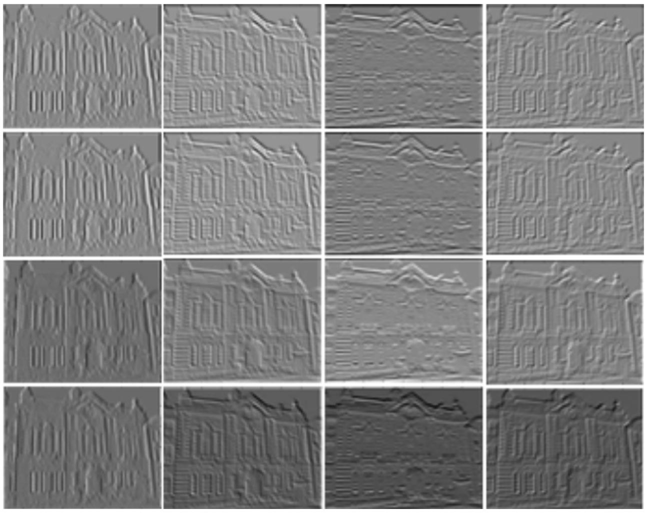


Fig. 3. Steerable filter responses at eight different orientations. The top two rows represent second-order Gaussian responses, and the bottom two rows correspond to their Hilbert transform responses.

B. Feature Pooling

From a biological perspective, a complex cell can be seen as polling input from an array of simple cells at different locations to generate its position-invariant response [5]. Due to this property, pooling has been widely utilized in feature extraction in many visual recognition algorithms. In SIFT [30], orientations are measured in the neighborhood of a keypoint, resulting in local orientation histograms; bag-of-features [35] use pooling for vector quantization of local descriptors.

Pooling consists of two mechanisms: sum pooling and max pooling. Sum pooling is a linear pooling mechanism which assigns equal weights to input and thus loses feature specificity. It cannot achieve size invariance. On the contrary, a max-like operation may present the cortical equivalence of the “windows of analysis” as in machine vision, where max pooling is such a nonlinear maximum operation as a key mechanism for object recognition in the cortex that corresponds to the truth that the largest receptive filed will always win. It selects the strongest (most active) response that corresponds to more robust response in the case of recognition in clutter. To this end, we use max pooling [5], [38] to achieve more compact representation and robustness to image noise by searching the max value of the steerable responses over local patches. This step preserves discriminative information while discarding irrelevance by combining feature responses over nearby regions into a more useful statistical representation that summarizes the joint distribution of the features over a local neighborhood. By carefully adjusting the pooling step of feature representation, recognition tasks can be facilitated. Max pooling is particularly well suited to separating very sparse features, and its efficiency has been demonstrated in [5], [38]. Here, each feature map is divided into 4×4 regions (as seen in Fig. 2), and each image is represented by a 256-dimensional feature vector.

C. Dimensionality Reduction

To alleviate computational complexity while preserving sufficient discriminative information in the subsequent recognition stage, the dimension of the feature vectors is reduced from 256 to 39 by LDA [33].

LDA is a supervised learning algorithm that takes class label information into account. Given a set of labelled training examples, it aims to separate examples from different classes far away while keeping those within the same class close to each other in the projected lower dimensional subspace.

Given that the original high-dimensional data points $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\}$ in $\mathbb{R}$ belong to c classes, the between example, and $m_i = (1/N_i) \sum_{j=1}^{n_i} \vec{x}_{i;j}$ is the mean value; $N = \sum_{i=1}^{c} N_i$ is the number of all training examples; and $m = (1/N) \sum_{i=1}^{c} \sum_{j=1}^{N_i} \vec{x}_{i;j}$ is the mean vector of the whole input data.

The formulation of LDA is

$$U_{opt} = \arg\max_{U} \frac{U^T S_b U}{U^T S_w U} \tag{17}$$

which finds the projection direction that maximizes the ratio between S_b and S_w in the projected low-dimensional subspace.

The generalized eigenvalue problem is $S_b U = \lambda S_w U$, and the resulted lower dimensional subspace is spanned by $U = \{\vec{u}_1, \vec{u}_2, ..., \vec{u}_L\} \cdot L \leq c - 1$. Herein, the covariance matrix of all training examples is $S_t = (1/N) \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\vec{x}_{i;j} - \vec{m}) (\vec{x}_{i;j} - \vec{m})^T = S_b + S_w$, which is also called the total-class scatter matrix.

IV. Experiments and Evaluation

Here, we first introduce the Sheffield Building Image Dataset (SBID) [27], where all experiments in this paper are conducted. Afterwards, we report the performance of SFBR and comparison results with two of the state-of-the-art building recognition systems.

A. Database Description

The SBID [27] makes the building recognition task more challenging by combining different variations together, e.g., rotation, scaling, different lighting conditions, viewpoint changes, occlusions, and vibration, including extremely highly variable lighting conditions and large viewpoint changes. This database consists of 3192 images taken from 40 buildings/categories, which include churches and a variety of old and modern buildings, such as exhibition halls and office buildings. The size of each image which ensures the computational efficiency and low memory requirement. A sample image for each category is given in Fig. 4.

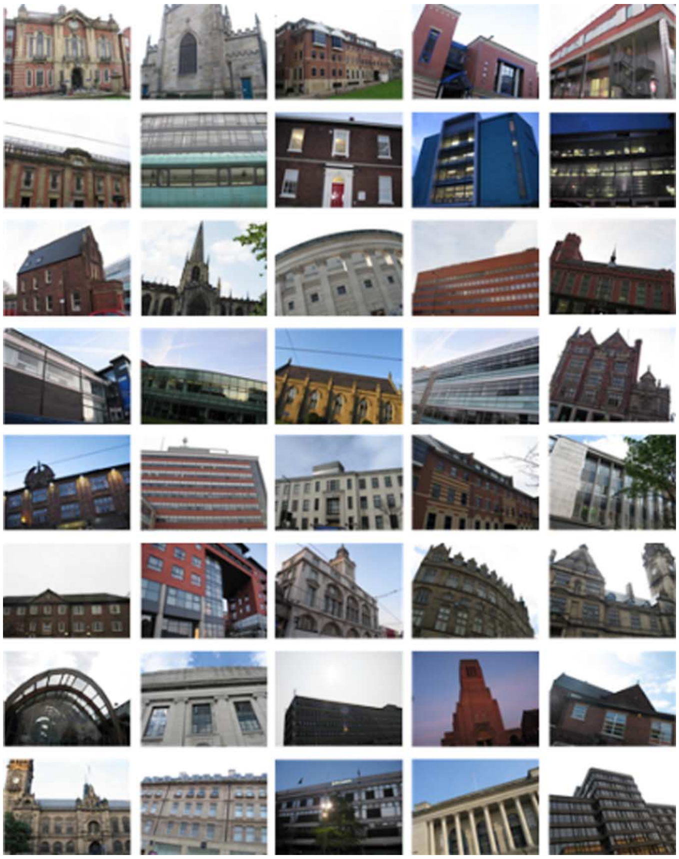


Fig. 4. SBID [27]. From left to right, top to bottom, each image represents a category from 1 to 40.

B. Performance of SFBR

Because of its good generalization ability and no requirement for prior knowledge about the data, we apply SVM [8], [46] for classification on SBID. SVMs are binary classifiers, which maximize the margin between two classes. They have been universally utilized as one of the most popular classifiers in various research areas, e.g., face recognition [17], texture classification [23], etc. Here, linear kernels $K(\vec{x}_i, \vec{x}_j) = \vec{x}_i^T \cdot \vec{x}_j$ are adopted because of its simplicity. The training process is as follows: we randomly select half of the images in each category for training and the rest are used for testing. This step is conducted for 20 times, and the average accuracy is calculated as 94.66%.

C. Comparison Results

To evaluate the performance of SFBR with that of the state-of-the-art building recognition systems, we compare the SFBR model with the hierarchical building recognition (HBR) system [49] and the biologically plausible building recognition (BPBR) scheme [27]. First, we simply introduce the mechanisms of these two systems as follows.

- HBR: The procedure of the HBR system is given as follows. Detected line segments are grouped into dominant vanishing directions and vanishing points are estimated by the expectation maximization (EM) algorithm: these two steps run simultaneously. Dominant vanishing directions here are referred to left, right, and vertical; vanishing points are defined as intersections of parallel lines. After

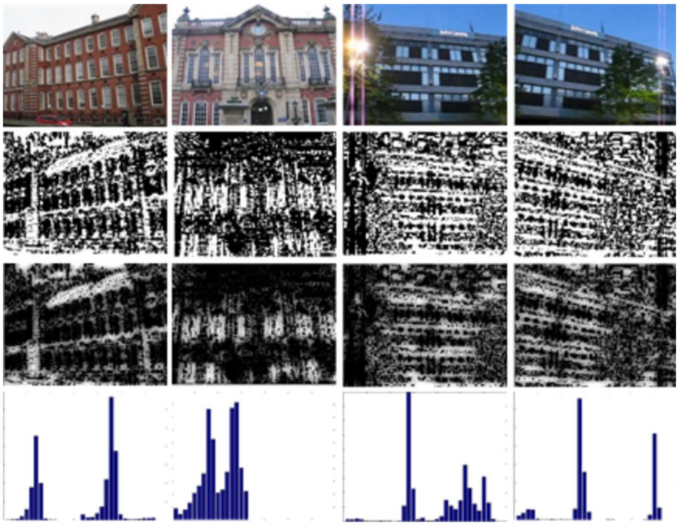


Fig. 5. Two views of building 1 and two views of building 38 with different viewing angles and lighting conditions. Top row: original images. Second row: pixel membership assigned using geometric constraints. Third row: pixel membership assigned after connected component analysis. Bottom row: indexing vector for each image.

that, image pixels with orientations aligned to the main vanishing directions, while satisfying some constraints defined by gradient magnitudes and directions, are divided into three groups, namely left, right, and vertical. Please note localized color histograms are only computed on these pixels. To reduce the sensitivity to viewpoint changes, pixels in the left group and right group are merged together, resulting in two groups in total. For each group, a single hue histogram, which is robust to illumination changes, is computed and quantized into 16 bins. Afterwards, the extracted two hue histograms are concatenated into a single vector.

The process is shown in Fig. 5: the first row shows the original images from Category 1 and Category 38 in the SBID; the second row presents the pixel membership assigned using geometric constraints; after connected component analysis, the refined pixel membership is given in the third row; the fourth row describes the extracted hue histograms of an image. Finally, histogram vectors are matched using the chi-square distance.

- BPBR: The BPBR scheme [27] is based on biologically inspired features that model the process of human visual perception. In BPBR, both saliency features [40] and gist features [36] were extracted to represent an image, where saliency features are obtained by extracting visual information from a raw image and gist features are acquired based on the extracted saliency features. To deal with the curse of dimensionality [3], the dimensionality of extracted features is reduced by LDA [33]. Afterwards, classification is conducted by the nearest neighbor rule [13] which assigns the same label to a test example as that of its nearest neighbor.

The comparison results of HBR, BPBR, and SFBR are given in Table I, where the newly proposed SFBR performs best with 94.66% as the accuracy, BPBR the second with an accuracy of 85.25%, and HBR the third with an accuracy of 73.32%.

TABLE I
COMPARISON RESULTS

Performance Evaluation	Building Recognition Algorithms		
	HBR	***BPBR***	***SFBR***
Average Precision (%)	73.32%	85.25%	94.66%

TABLE II
COMPARISON RESULTS

Performance Evaluation	Dimensionality Reduction Algorithms in SFBR		
	DLA	***LPP***	***LDA***
Average Precision (%)	88.43%	93.39%	94.66%

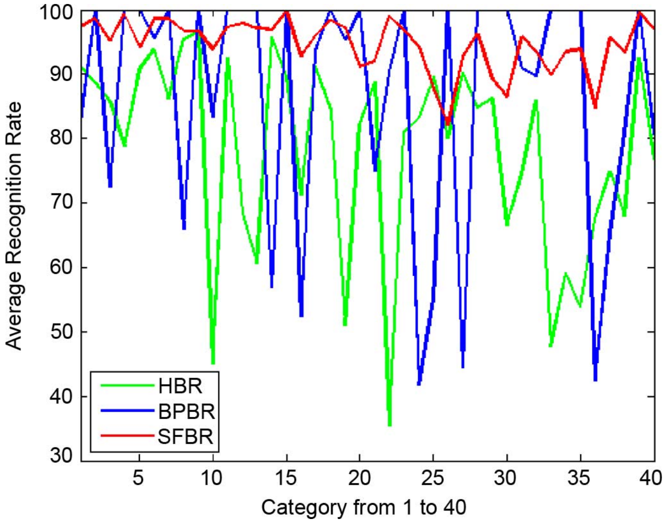


Fig. 6. Comparison results of HBR, BPBR, and SFBR for each category in SBID [27], [51].

As a further study to evaluate the effectiveness of different dimensionality reduction methods, in SFBR, we substitute LDA by two of representative manifold learning algorithms, i.e., locality preserving projections (LPP) [19] and discriminative locality alignment (DLA) [48]. The comparison results of DLA, LPP, and LDA are reported in Table II, where the performance of LPP is 93.39% and the performance of DLA is 88.43%; LDA performs best for our experimental work.

The recognition performance of HBR, BPBR, and SFBR for different individual categories is shown in Fig. 6. As we can see, for most categories in the SBID, the performance of the SFBR model is the most stable and better than the others. HBR works well for Categories 8 and 9. However, its performance is poor for some categories, e.g., Categories 10 and 22. This indicates the HBR algorithm performs well for buildings with modest challenges, but it cannot deal with large illumination variations and viewpoint changes. Why does SFBR outperform HBR and BPBR for building recognition? Different features contribute to different strengths in mimicking perceptual saliency. For a building recognition task, edge information is the most important feature in discriminating a building from another since each building contains windows, doors, and bricks, for example. In

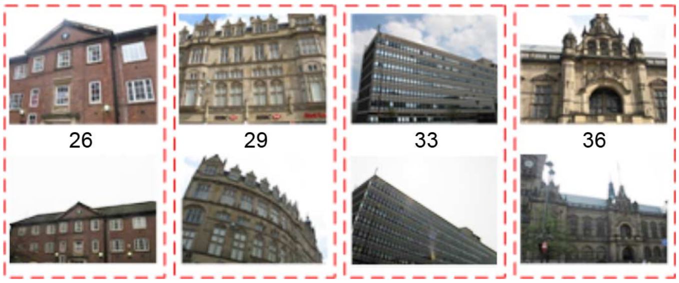


Fig. 7. Visual examples with relatively poor recognition rates (less than 90%) in Categories 26, 29, 33, and 36, respectively.

SFBR, we use second-order steerable filters, which are more suitable for extracting edge information and enhancing images. In addition, steerable filters are orientation-selective, which are able to deal with edges at arbitrary orientations and so have potential for the task of building recognition. On the other hand, neither HBR nor BPBR includes both characteristics mentioned above. Another reason for SFBR performing better than BPBR is that BPBR simply sums up the activities of units (i.e., sum pooling), while SFBR adopts max pooling, which implies that the largest receptive filed will always win.

A further point worth noting is that although the average performance of SFBR is 94.66%, the recognition rates in some categories are less than 90% because of the immense challenges in the SBID. Visual examples with relatively poor recognition rates are shown in Fig. 7. As we can see, Categories 26 (82.05%) and 33 (89.83%) have large illumination changes, while the pictures in Categories 29 (89.19%) and 36 (84.76%) were taken from a wide range of viewing angles and some parts of the buildings are missing.

V. Conclusion

In this paper, we present a novel building recognition model, which combines local oriented features, feature pooling, and dimensionality reduction for computational effectiveness and efficiency. Based on experimental comparisons with two state-of-the-art building recognition systems, this practical, modular, and easy-to-implement model is very effective for building recognition. However, it is important to make clear that this model is not meant to substitute building recognition algorithms. Rather, it might be regarded as an alternative solution for many vision-based applications. Although the suggested second-order steerable filters were applied here to a building recognition task, they can be easily extended to fourth-order for other computer vision and pattern recognition applications.

References

[1] H. Ali, G. Paar, and L. Paletta, “Semantic indexing for visual recognition of buildings,” in *Proc Int. Symp. Mobile Mapping Technol.*, 2007, pp. 28–31.

[2] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” *Adv. Neural Inform. Process. Syst.*, vol. 14, pp. 585–591, 2002.

[3] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ, USA: Princeton Univ., 1961.

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

[5] Y.-L. Boureau, J. Ponce, and Y. Lecun, “A theoretical analysis of feature pooling in visual recognition,” in *Proc. Int. Conf. Mach. Learning*, 2010, pp. 111–118.

[6] L. Breiman, “Bagging predictors,” *Mach. Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[7] D. Bruckner, C. Picus, R. Velik, W. Herzner, and G. Zucker, “Hierarchical semantic processing architecture for smart sensors in surveillance networks,” *IEEE Trans. Ind. Inf.*, vol. 8, no. 2, pp. 291–301, May 2012.

[8] J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[9] X. Cao, B. Ning, P. Yan, and X. Li, “Selecting key poses on manifold for pairwise action recognition,” *IEEE Trans. Ind. Inf.*, vol. 8, no. 1, pp. 168–177, Feb. 2012.

[10] Y.-C. Chen and J.-H. Chou, “Mobile robot localization by RFID method,” in *Proc. Int. Conf. Digital Telecommun.*, 2012, pp. 33–38.

[11] S. Chen, J. Zhang, Y. Li, and J. Zhang, “A hierarchical model incorporating segmented regions and pixel descriptors for video background subtraction,” *IEEE Trans. Ind. Inf.*, vol. 8, no. 1, pp. 118–127, Feb. 2012.

[12] Y.-C. Chung, T. X. Han, and Z. He, “Building recognition using sketch-based representations and spectral graph matching,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2014–2020.

[13] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

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

[15] N. Groeneweg, B. Groot, A. Halma, B. Quiroga, M. Tromp, and F. Groen, “A fast offline building recognition application on a mobile telephone,” *Adv. Concepts Intell. Vis. Syst.*, pp. 1122–1132, 2006.

[16] A. Jain and A. Vailaya, *Image Retrieval Using Color and Shape Pattern Recogn.*, vol. 29, no. 8, pp. 1233–1244, 1996.

[17] G. Guo, S. Z. Li, and K. L. Chan, “Support vector machines for face recognition,” *Image Vis. Comput.*, vol. 19, no. 9–10, pp. 631–638, 2001.

[18] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Proc. Alvey Vis. Conf.*, 1988, pp. 147–151.

[19] X. He and P. Niyogi, “Locality preserving projections,” in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2003.

[20] R. Hutchings and W. Mayol-Cuevas, “Building recognition for mobile devices: Incorporating positional information with visual features,” Comput. Sci., Univ. Bristol, Bristol, U.K., Tech. Rep. CSTR-06-017, 2005.

[21] Q. Iqbal and J. K. Aggarwall, “Applying perceptual grouping to content-based image retrieval: Building images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 1999, vol. 1, pp. 42–48.

[22] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Berlin, Germany: Springer, 2002.

[23] K. I. Kim, K. Jung, S. H. Park, and H. J. Kim, “Support vector machine for texture classification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1542–1550, Nov. 2002.

[24] T. S. Lee, “Image representation using 2D gabor wavelets,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, Oct. 2003.

[25] S. Lee and J.-B. Song, “Mobile robot localization using range sensors : consecutive scanning and cooperative scanning,” *Intl J. Control, Autom., Syst.*, vol. 3, no. 1, pp. 1–14, 2005.

[26] J. Li and N. M. Allinson, “A comprehensive review of current local features for computer vision,” *Neurocomputing*, vol. 71, no. 10-12, pp. 1,771–1,787, 2008.

[27] J. Li and N. M. Allinson, “Subspace learning-based dimensionality reduction in building recognition,” *Neurocomputing*, vol. 73, no. 1–3, pp. 324–330, 2009.

[28] Y. Li and L. G. Shapiro, “Consistent line clusters for building recognition in CBIR,” in *Proc. IEEE Int. Conf. Pattern Recogn.*, 2002, vol. 3, pp. 952–956.

[29] H. Liu, “A fuzzy qualitative framework for connecting robot qualitative and quantitative representations,” *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 6, pp. 1522–1530, Dec. 2008.

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

[31] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

[32] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," in *Proc. Brit. Mach. Vision Conf.*, 2002, pp. 384–393.

[33] G. J. McLachlan*, Discriminant Analysis and Statistical Pattern Recognition*. New York, NY, USA: Wiley-Interscience, 1992.

[34] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1,615–1,630, Oct. 2005.

[35] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2006, vol. 3954, pp. 490–503.

[36] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Intl Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[37] J. Park and J. Lee, "A beacon color code scheduling for the localization of multiple robots," *IEEE Trans. Ind. Inf.*, vol. 7, no. 3, pp. 467–475, Aug. 2011.

[38] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[39] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2,323–2,326, 2000.

[40] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[41] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cog. Psychol.*, vol. 12, pp. 97–137, 1980.

[42] H.-H. Trinh, D.-N. Kim, and K.-H. Jo, "Facet-based multiple building analysis for robot intelligence," *J. Appl. Math. Computation*, vol. 205, no. 2, pp. 537–549, 2008.

[43] R. S. Turner*, In the Eye's Mind: Vision and the Helmholtz Hering Controversy*. Princeton, NJ, USA: Princeton Univ., 1994.

[44] T. Tuytelaars and L. V. Gool, "Matching widely separated views based on affine invariant regions," *Int. J. Comput. Vis.*, vol. 59, no. 1, pp. 61–85, 2004.

[45] M. M. Ullah, A. Pronobis, B. Caputo, J. Luo, R. Jensfelt, and H. I. Christensen, "Towards robust place recognition for robot localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2008, pp. 530–537.

[46] V. Vapnik*, The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag, 1995.

[47] Q. Zhang, L. Lapierre, and X. Xiang, "Distributed control of coordinated path tracking for networked nonholonomic mobile vehicles," *IEEE Trans. Ind. Inf.*, vol. 9, no. 1, pp. 472–484, Feb. 2013.

[48] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2008.

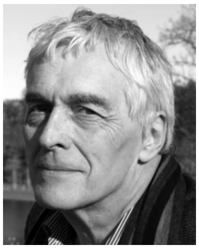
[49] W. Zhang and J. Košecká, "Hierarchical building recognition," *Image Vis. Comput.*, vol. 25, no. 5, pp. 704–716, 2007.

[50] M. A. Zmuda, A. Elesev, and Y. T. Morton, "Robot localization using RF and inertial sensors," in *Proc. IEEE Nat. Aerosp. Electron. Conf.*, 2008, pp. 343–348.

[51] 2002 [Online]. Available: http://www.ece.osu.edu/~maj/osu_svm/

Jing Li received the Ph.D. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., in 2012.

She is currently a Lecturer with the School of Information Engineering, Nanchang University, Nanching, China. Previously, she was a Research Associate with the University of Sheffield. She has authored or coauthored in various peer-reviewed journals.



Nigel Allinson (SM'12) holds the Distinguished Chair of Image Engineering with the University of Lincoln, Lincoln, U.K. He previously held Chairs with the University of Sheffield and UMIST (Manchester) and was awarded a Member of the British Empire medal in H.M. Queen's Birthday Honours (2012) for services to engineering. He has authored or coauthored over 300 scientific papers and has cofounded five spinoff companies based on his research.