Information Sciences 277 (2014) 406–420

Contents lists available at ScienceDirect

Information Sciences

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

Building recognition in urban environments: A survey of state-of-the-art and future challenges

Jing Li [a], Wei Huang [a], Ling Shao [b,*], Nigel Allinson [c]

[a] School of Information Engineering, Nanchang University, Nanchang 330031, China
[b] Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 3JD, UK
[c] School of Computer Science, University of Lincoln, Lincoln LN6 7TS, UK

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form 2 January 2014
Accepted 15 February 2014
Available online 26 February 2014

Keywords:
Building recognition
Wide baseline matching
Dimensionality reduction
SIFT
Robot localization
Visual navigation

ABSTRACT

Building recognition in urban environments aims to identify different buildings in a large-scale image dataset. This identification facilitates the annotation of any visual object to a building's façade and is an essential step in a variety of applications, such as automatic target detection in surveillance, real-time robot localization and visual navigation, architectural design, and 3D city reconstruction. Because of its importance, a significant number of building recognition systems have been proposed in recent years. Nevertheless, there is no systematic survey of building recognition in urban environments yet. To this end, we present a comprehensive review of the dominant building recognition systems by first grouping them into two categories: (i) effectiveness approaches that mainly focus on the improvement of recognition performance and (ii) efficiency methods that attempt to enhance the recognition speed. Effectiveness approaches are further categorized into two different groups: (i) feature representation-based algorithms and (ii) wide baseline matching-based methods. Efficiency methods are divided into: (i) dimensionality reduction-based methods and (ii) clustering-based algorithms. We provide analysis and discussions on each type of method and summarize their advantages and weaknesses in depth. Furthermore, we outline future research directions and associated challenges in this promising area. This survey can serve as a starting point for new researchers in building recognition to generate new ideas according to their specific requirements.

1. Introduction

Building recognition in urban environments aims to distinguish between different buildings in a large-scale image dataset. The identification of buildings can contribute to various applications, such as automatic target detection in surveillance, real-time robot localization [47] and visual navigation [14,30], architectural design, and 3D city reconstruction [1,36,41,50], as shown in Fig. 1. Building recognition enables robots to easily and accurately localize themselves in outdoor scenes and therefore is helpful for robot localization and navigation; in traffic monitoring, moving vehicles are usually detected under the circumstances in urban area; building recognition allows building designers or developers to efficiently search the most desirable building models from the Internet or databases when designing a new building; building recognition also plays a fundamental part in 3D city reconstruction for urban planning or augmented reality in video games. Having become such an

* Corresponding author. Tel.: +44 1142225841.
E-mail address: ling.shao@sheffield.ac.uk (L. Shao).

http://dx.doi.org/10.1016/j.ins.2014.02.112
0020-0255/

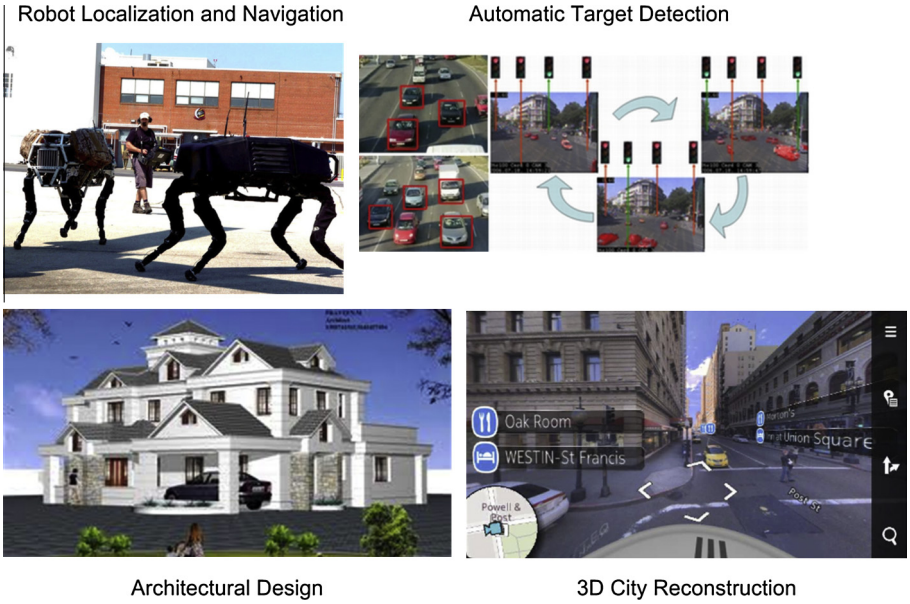

Fig. 1. Different applications of building recognition. These pictures are taken from [59–62], respectively.

important computer vision task, building recognition has attracted considerable attention from both aerial images [55] and city building images.

Building images can be separated into two types: (i) aerial images and (ii) city building images. These two kinds of images are fundamentally different in appearance, as shown in Fig. 2. Aerial images of a particular scene comprise a number of overlapping digital photographs taken by different video cameras or sensors fitted on a plane or other high-altitude platform that flies back and forth above a city. The images are usually of high-resolution and can be affected by the speed and altitude of the aircraft, weather conditions, the type of data, and the post-processing methods applied. Compared with low-resolution imagery, ultra-high resolution datasets of aerial images require enormous amounts of storage space and computational costs for building recognition. City building images are usually collected by taking pictures with a digital camera/video camera from a street-level view and from different viewing angles. Mayer [31] presented a review on building recognition approaches applied to aerial images; nevertheless, there is no survey paper about city building images. To this end, this paper surveys the state-of-the-art building recognition techniques on city building images from different aspects, covering a variety of knowledge in the computer vision research field, especially on effective features for building recognition.

Building recognition is usually deemed to be an object recognition or a content-based image retrieval problem [21,27] for a specific category. Compared to general object recognition tasks [51–54] on both images and videos, city building recognition is more challenging because most building images contain both human-made objects (e.g., walls, doors, and windows) and natural scenes (e.g., trees), and they often exhibit repetitions and planar surfaces. Moreover, images taken from the same building could demonstrate a wide range 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. Therefore, an ideal building recognition technique should be sensitive enough to identify an individual building while robust to different geometric and photometric image transformations (e.g., rotation, scaling, viewpoint changes, and different lighting

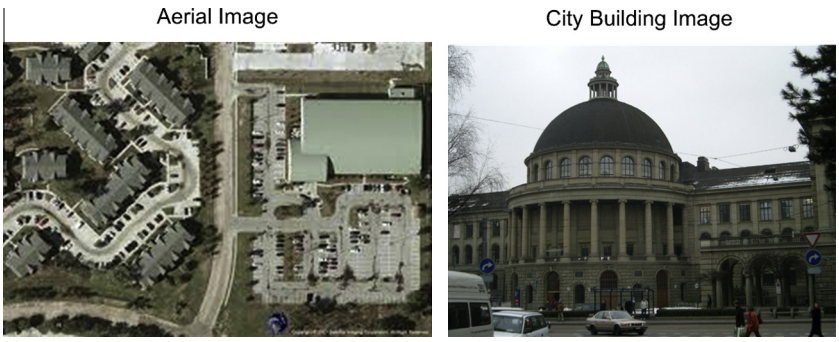

Fig. 2. Examples of an aerial image and a city building image, respectively.

conditions). Furthermore, it should be efficient enough for real-time applications in large-scale databases. These characteristics of building recognition have led to research on the following aspects: (1) feature representations of visual models that can accurately describe buildings and are robust to small image transformations and (2) fast indexing that can improve an algorithm's efficiency while alleviating data storage space and computational complexity.

Generally, a building recognition system consists of three parts: (i) feature representation; (ii) feature matching; and (iii) classification. Feature representation usually serves as the first step in the whole building recognition process and directly affects the subsequent recognition performance. It describes different objects in an image by extracting either global features [26,45,22,19,28] or local features [29,23,33,23,2,39], which are considered as two types of complementary features. Global features are extracted from all pixels in a whole image; local features, referring to image patterns (e.g., a local image region or an object of interest) that are different from those in their neighborhoods, could describe local information of an image and tend to be invariant to both geometric and photometric image transformations. To identify buildings in the building recognition task, global features refer to a building's color, shape, texture, etc.; while local features can be deemed as salient points and descriptors calculated on the neighborhoods of them for distinguishing different parts of a building, e.g., walls, doors, windows, etc. After feature representation for every image, feature matching is conducted to find the correspondences between a pair of building images, i.e., the query image and a reference image in the database via distance metrics (e.g., Euclidean or Mahalanobis distance). Finally, classification is conducted to determine the best match, where classifiers [11,48] over statistical models combine the outputs of either global or local appearance feature vectors to maximize the quality of the output on a training set. Recognition performance is usually evaluated by precision and recall, where precision is defined as the percentage of the relevant images in top retrieved images and recall is defined as the percentage of the relevant images in all positive images in the database.

A number of building recognition systems [34,42,24,45,14,18,22,15,25,26,25,19,28,57,34,58,10,2] have been proposed in the last few years. In this survey, we first classify them into two categories: (i) effectiveness approaches that mainly focus on the improvement of recognition performance and (ii) efficiency methods that speed up the recognition system. Effectiveness approaches can be further categorized into two different groups: (i) feature representation-based algorithms and (ii) wide baseline matching-based methods. Efficiency methods are divided into: (i) dimensionality reduction-based methods and (ii) clustering-based algorithms.

The objective of this survey is to introduce representative building recognition techniques on city building images and give researchers who are interested in building recognition an overview of the state-of-the-art methods. The remainder of this survey is as follows. In Section 2, we describe current city building datasets and their complexity. In Section 3, we review representative building recognition approaches in terms of their effectiveness and efficiency. Finally, we summarize the paper and suggest promising future directions of building recognition from a wider variety of real-world data in Section 4.

2. Datasets and their complexity

To date, most of the building recognition techniques have been evaluated on the following datasets: (i) Zurich building database (ZuBuD) [37]; (ii) Sheffield Building Image Dataset (SBID) [26]; and (iii) Oxford Buildings Dataset (OBD) [34]. The complexity of different datasets is mainly due to the image contents and resolutions. In the following sub-sections, these datasets and their associated complexity will be introduced. Furthermore, we point out the challenges of the building recognition task for various types of data, such as surveillance videos, Google Street View images, and satellite images.

2.1. Zurich building database

Zurich Building Database (ZuBuD) [37], generated in 2003, contains 201 buildings of Zurich City. Each building was viewed from five arbitrary angles, resulting in 1005 images in total. The size of each image is fixed at 640×480. Sample images of three different buildings are given in Fig. 3. As we can see, the images in ZuBuD lack large viewpoint and illumination changes, and this dataset does not consider the combination of different types of variance. What is more, query images of ZuBuD are too easy to differentiate between simple methods and advanced algorithms [15]. As a result, although ZuBuD has been widely used in the literature [38,15,57], it is insufficient to serve as a good benchmark for evaluating the performance of different building recognition algorithms.

2.2. Sheffield building image dataset

Sheffield Building Image Dataset (SBID) [26] makes the building recognition task more challenging by combining different types of variations together, including highly variable lighting conditions and large viewpoint changes. As we can see in Fig. 4, the images in SBID possess various challenges, i.e., rotation, scaling, different lighting conditions, viewpoint changes, occlusions, and vibration – even humans cannot easily tell whether the images in the second row are from the same building or not. SBID consists of 3192 images taken from 40 buildings, which include churches and a variety of modern buildings, such as exhibition halls and office buildings. Still images and video clips were taken around the University of Sheffield and the Sheffield City centre at different times on separate days in 2008, where different times cover early morning, noon,

Fig. 3. Example images of the ZuBuD database. Each row shows images from the same building.

Fig. 4. Sample images of the SBID dataset. These three rows show sample images from Categories 1, 10, and 38, respectively.

mid-afternoon, and early evening. Still images for each building were taken from different viewpoints, varying from three to nine views; video clips were also obtained with multiple views by moving the camera from side to side. Furthermore, some videos were captured by walking from one side to another, which creates the challenge of movement/camera-shake. The size of both still images and sampling frames of video clips is fixed to 160×120 in order to ensure computational efficiency and low memory requirements.

2.3. Oxford buildings dataset

Oxford Buildings Dataset (OBD) [34] consists of a variety of real-world images from Internet photo collections, which were taken from different kinds of cameras under varying lighting conditions. It is composed of 5063 high-resolution images of 17 different Oxford landmarks, i.e., the size of each image is either 1024×768 or 768×1024. Sample images are shown in Fig. 5. Different from above-mentioned databases, images in OBD were obtained by searching query keywords of Oxford landmarks (such as "Oxford All Souls" and "Oxford Christ Church") from Flickr [63] – one of the largest Internet photo-sharing websites. Therefore, not only building images were collected, but also distractors, i.e., images containing other objects were collected when searching query keywords. For the evaluation of object retrieval systems, OBD has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each of which is represented by 5 possible queries. Each image and landmark in the dataset is assigned to one of the following labels: (i) Good; (ii) OK; (iii) Junk; and (iv) Absent, depending on the quality of images and whether the object/building is present or not.

Fig. 5. Sample images of the OBD dataset. Each row shows sample images collected by searching "All Souls Oxford", "Ashmolean Oxford", and "Oxford" as queries, respectively.

Systems relying on photos captured with the same calibrated camera at a regular sampling rate may not perform well because images collected from the web have none of these simplified characteristics. Consequently, OBD is a challenging dataset for evaluating the effects of different visual vocabularies and spatial ranking in object retrieval and can distinguish simple algorithms from robust ones. However, photographs from the Internet are inherently redundant since many of them were taken from nearby viewpoints. Besides, it is not easy to recover 3D geometry from this large collection of images.

However, the above-mentioned datasets contain only images. Although SBID captured some videos, it sampled these video clips into frames, which also fall into the image type. Except for images stored in databases, building recognition may be directly conducted on various types of data (e.g., surveillance videos, Google Street View (GSV) images, and satellite images): each with different challenges. For surveillance videos in cities, lower quality videos introduce more challenges for the building recognition task. Moreover, desirable building recognition techniques should be especially robust to illumination changes and occlusions in crowded scenarios. For GSV images, one challenge is how to efficiently collect GSV images we need from the web; the other challenge is that the images may not contain the targets (buildings) or they are taken from very large viewing angles. Satellite imagery entails similar challenges with aerial images: the images are usually of high-resolution which requires large amounts of storage space and computational costs. In summary, different data types may bring varying degrees of challenges for building recognition.

3. Existing techniques for building recognition

In this section, we separate existing building recognition techniques into two main categories: (1) effectiveness approaches and (2) efficiency approaches. Based on the procedure of building recognition, we roughly divide effectiveness approaches into two categories: (i) feature representation-based algorithms; (ii) wide baseline matching-based approaches; and (iii) others. Efficiency approaches are further categorized into three classes: (i) dimensionality reduction-based methods; (ii) clustering-based algorithms; and (iii) others. In the following sub-sections, we give a brief overview of the underlying algorithms.

3.1. Effectiveness approaches for building recognition

Effectiveness approaches are divided into two categories: (i) feature representation-based algorithms [34,42,24,45] and (ii) wide baseline matching-based methods [14,18,22]. Feature representation-based algorithms focus on the process of feature extraction in building recognition; wide baseline matching-based approaches identify corresponding building facades from two different views by efficiently matching corresponding feature points between a query image and a reference image. Representative algorithms for these two categories will be introduced in separate sub-sections.

3.1.1. Feature representation-based algorithms

In [38], an indexing method, called hyper-polyhedron with adaptive threshold (HPAT), was proposed to reduce the number of feature vectors in searching for the nearest neighbors. It approximates the hyper-sphere with a hyper-polyhedron rather than a hyper-cube, where the illustrations of the hyper-cube and hyper-polyhedron in 2D space and 3D space are given in Figs. 6 and 7, respectively. As we can see, the hype-cube approximation includes more useless corner points (the corner p_c falls within the hyper-cube, but not within the hyper-sphere), which results in more search space than the hyper-polyhedron approximation. To be more precise, the volume of the hype-cube is $(2r)^n$ given r as the radius of the hype-sphere; while the volume of the hyper-polyhedron is $(2r)^n(\sqrt{2}-1)^n\left[1+\sum_k^n\binom{n}{k}2^{1-k/2}\right]$ which grows slowly with the increasing number of dimensions and keeps the volume of the hyper-sphere relatively constant.

The procedure of building recognition is carried out as follows. Intensity-based regions [46] are extracted at multi-scale intensity extrema of a Gaussian scale space, and then each region is described by a set of nine generalized color moment invariants. The extracted local features are robust to illumination and viewpoint changes. Nevertheless, they lead to feature matching in a higher dimensional space compared with global features. To this end, HPAT can improve the efficiency in localizing the nearest neighbors in the feature space and so reduce computational time, where the Mahalanobis distance is utilized by taking into account the differences and correlation among the elements of a feature vector. For building recognition, this model was tested on ZuBuD [37] and the recognition rate was 77.3%. It can identify pictures of the same building from a wide range of viewpoints in a large image database. However, this model embeds the following shortcomings: (1) computational time is mainly spent on the extraction of invariant regions, which indicates simple and effective features can alleviate the whole computational cost; (2) ZuBuD is a relatively small building image dataset, so the model may not work well for larger databases since recognition rates decline due to too many similar regions found in other images, especially for building images that have similar colors; and (3) it does not determine the query pose.

Motivated by the robustness of local features to different geometric and photometric transformations, Li and Allinson [24] proposed the steerable filter-based building recognition (SFBR) model which is able to select oriented features with arbitrary orientations and thus can deal with edge information with varying angles. To achieve invariance to small shifts in position and changes in lighting conditions, max-pooling [44,8] – a key mechanism for object recognition in the cortex, is used to preserve discriminative information by searching the max value of the steerable responses over local patches. Afterwards, linear discriminant analysis (LDA) [32] is applied to reduce the dimensionality of feature vectors via projecting data points into a lower subspace. Finally, a support vector machine (SVM) [48], maximizing the margin between positive examples and negative examples, is used to discriminate different buildings because of its good generalization abilities and no requirement for prior knowledge about the data. The SFBR model demonstrates the promising properties and capabilities of local features for describing the characteristics of the components of a building (e.g., windows, doors, and bricks). Although it is simple, it offers a modular, computationally efficient, and effective alternative to other building recognition techniques.

Suleiman et al. [42] utilized the SIFT operator [29] as an image texture descriptor to identify building façades and estimate the calibrated camera geolocation (i.e., absolute camera position and orientation) in building images. The aim is to

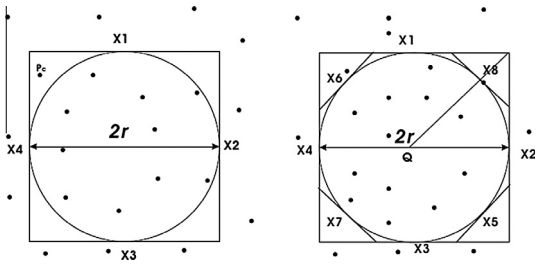

Fig. 6. Hyper-cube and hyper-polyhedron in two dimensions, the corner p_c falls within the hyper-cube, but not within the hyper-sphere. The images are from [38].

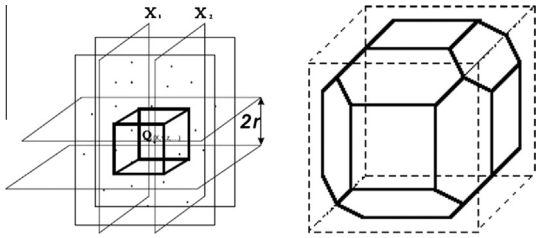

Fig. 7. Reduction of the volume from hyper-cube approximation to hyper-polyhedron approximation by using more planes – illustration in 3D space. The images are from [38].

enhance and complement a registered geographic information systems (GIS) database of 70 façade images taken from 20 buildings in the city center of Saint Etienne. The procedure of this method contains the following steps:

(1) automatically crop the whole building façade for each image in the database. For non-accurate cropped images, the real boundaries are manually extracted;
(2) extract SIFT image features for each cropped façade;
(3) find correspondences between the query image and the reference facades in the 3D GIS database by calculating the distances between their descriptors;
(4) eliminate false matches by calculating the homography constraints by direct linear transformation (DLT) and minimizing the geometrical distance for all the 2D/3D correspondences using a nonlinear iterative Levenberg–Marquardt approximation; and
(5) add SIFT descriptors with the 3D positions of the interest points and four corner points of each cropped building façade to the GIS database.

The 3D positions of the 2D interest points and four corners of the query image can be determined in this way and this can be used as an initialization phase in the automatic registration process. Nevertheless, the estimation of camera position and orientation depends on affine rotation and the performance drops significantly if the angle between the discovered facade and the image projection plane is larger than 40°. Therefore, the estimation using multiple non-coplanar facades is rather bad when the rotation angle is large in most urban pictures. Moreover, it does not perform well for a building that has similar textures or a building with a glass facade that could reflect the facades of other buildings.

Zhang et al. [56] recognized building architecture styles based on their morphological characteristics captured by highly discriminative blocklets, which represent basic architecture components as well as their spatial arrangements and are quantized by a hierarchical sparse coding method. The effectiveness of this approach was validated over 10,000 buildings from nine categories of architecture styles, and experimental results demonstrate that the approach outperforms some specific building/place recognition models.

While most of the above-mentioned building recognition algorithms focus on single-building recognition, i.e., each image only contains one dominant building, Trinh et al. [45] proposed a method to recognize multiple buildings in the image database of Ulsan metropolitan city in South Korea. Since the main part of a building consists of windows, doors, and walls, facets of each building are extracted based on line segments and vanishing point detection. Afterwards, wall color histograms are first computed on the pixels satisfying some constraints for selecting candidate models that are robust for the multi-building recognition task or a single building containing several faces. SIFT features [29] are utilized to describe each building while only those detected keypoints with scales above 2 are further represented by the SIFT descriptor. For each test image, its closest model is selected according to the nearest neighbor rule. Basically, multi-building recognition is more challenging than that only considers a single building in an image. Nevertheless, by testing the proposed method for building recognition and segmentation of building images from non-building images, this algorithm is claimed to outperform all other approaches and serves as the only current method that is able to recognize multiple buildings in an image.

3.1.2. Wide baseline matching-based approaches

In building recognition, wide baseline matching [14,30,46] identifiers corresponding building facades from two different views by efficiently finding corresponding feature points between a query image and a reference image that were taken with significant viewpoint changes and under different lighting conditions. They follow the basic scheme of local feature representation which is briefly described as follows. Firstly, interest points are detected by corner detectors or region detectors. Afterwards, a local region is constructed around each interest point, in such a way as to adapt its shape to the viewing angle and keep the part of the scene it encloses fixed. All regions are then described by a local descriptor, and finally, the best match is determined using a voting scheme based on a distance measure between calculated descriptors. Fig. 8 gives an exemplar of wide baseline matching between two views, where SIFT features [29] were extracted and the correspondences were matched by the homography that maps pixels between two views of the same plane.

Fig. 8. An exemplar of wide baseline matching between two views.

A fast wide baseline matching algorithm [14], which allows for fast matching for the localization and recognition of natural landmarks, was developed for semi-automatic visual navigation. In every image, affine invariant column segments, i.e., vertical columns of pixels between a consecutive pair of local gradient maxima, are extracted. For each column segment, a descriptor vector is computed based on geometrical, color and intensity information. Afterwards, the Mahalanobis distance between the descriptor vectors and the horizontal distance between line segments are utilized for matching. Because of repetitive elements in an image, the column segments are grouped into clusters, each of which is represented by a prototype column segment associated with the average descriptor vector. In support of fast matching, a kd-tree of the reference image data is built of the cluster prototypes. Finally, RANSAC is applied to filter out mismatches.

Hutchings and Mayol-Cuevas [18] designed a building recognition system for mobile devices by locating a building from its position in world space. Given a query image, its local features are first extracted by the Harris corner detector [17] and then described by the SIFT descriptor [29]. Afterwards, they are matched with extracted features of every reference image in the database. Since building images are usually taken from different distances, scaling is not a trivial issue. To ensure there are enough matches between two images from the same building (even if the building is far away from the camera), a scale is selected for each query image by utilizing its GPS position in the matching process. This results in the reduction of search space and computational cost. To cope with viewpoint changes, homography that maps pixels between two views of the same plane is estimated by RANSAC and the structural resemblance is measured by the angle error of the spatial arrangement of matches. As seen in Fig. 9, p is a pixel of a view from camera A and q is a pixel of a view from camera B. The relation between them can be modeled by $p = Hq$ with $H = K(R - (tn^T)/d)K^{-1}$ being the homography matrix calculated by rotating and translating camera B with respect to A, where R is the rotation matrix, t is the translation, K is the camera calibration matrix, n is the plane normal, and d is the distance to the view from A. Most outliers can be removed with these steps, but the system fails in dealing with very large viewpoint changes.

For augmented reality-based navigation systems, Kim et al. [22] detected building locations based on edge and block information from video clips captured in Dajeon City (Korea) by a camera equipped in a moving vehicle. The process is conducted as follows:

(1) detect edges for each image frame;
(2) use the predefined mask to filter out vehicles and road areas;
(3) divide each image frame into small-sized blocks;
(4) for each block, calculate slopes and length of the edge segments in horizontal and vertical directions respectively;
(5) determine the search region by detecting and removing trees and background regions based on the calculated slopes in Step 5; and
(6) determine the building area and recognize buildings by performing block-based edge tracing in adjacent blocks.

This building recognition technique is performed only within the search region, which significantly reduces the processing time and makes the algorithm focus on highly possible areas in an image frame. The algorithm was tested on 42 video clips and the overall recognition performance is 88.9%. However, the detection rate for clips with more complex environments is only 62.3% since buildings in these image frames may be overlapped by other objects, e.g., trees, traffic signals, and street lights. This means the algorithm cannot deal with occlusions.

3.1.3. Summary

Most effectiveness approaches for building recognition are based on local features that are not only very effective to capture discriminative information for building recognition, but also invariant to geometric and photometric changes. The differences among these approaches lie in the way that interest point detectors, local invariant regions, and local descriptors

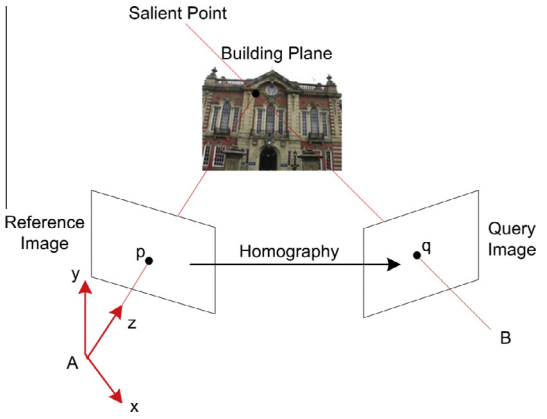

Fig. 9. Two views of the same building are related directly by a homography.

Table 1
Recognition rates reported by various effectiveness approaches on different databases.

Approach	Database	Performance (%)
HPAT indexing [38]	ZuBuD	77.3
SFBR [24]	SBID	94.7
Multiple buildings [45]	Building images in Ulsan metropolitan city	97.5
Fast wide baseline matching [14]	ZuBuD	92.0
Kim's method [22]	Video clips of Dajeon City (Korea)	95.9

are extracted. The individual recognition rates of various effectiveness approaches are provided in Table 1. Since the methods were not tested on the same building image dataset, it is not sensible to directly compare their performance only based on the recognition rates.

3.2. Efficiency approaches for building recognition

A series of algorithms were proposed for fast processing. Efficiency approaches can be categorized into three classes: (i) dimensionality reduction-based methods [15,25,26]; (ii) clustering-based algorithms [19,28,57,34,58,10]; and (iii) others [2]. Dimensionality reduction-based methods improve the recognition efficiency by alleviating the dimensionality of feature vectors and clustering-based algorithms aim to discover the relationships among different image structures by grouping them into different clusters.

3.2.1. Dimensionality reduction-based methods

Most building recognition systems suffer from the curse of dimensionality [5] due to the high dimensions of extracted feature vectors. In order to eliminate feature redundancy and make the data more compact, dimensionality reduction (DR) [16] finds a projection to reduce the original higher dimensional feature space to a much lower dimensional subspace, thereby alleviating the computational costs for the subsequent recognition process. DR approaches can be mainly divided into linear subspace methods (LSMs) and manifold learning algorithms. LSMs project the original higher dimensional data points $\vec{x}_i \in \Re^n$ $(1 \leqslant i \leqslant N)$ into a lower dimensional space $\vec{y}_i \in \Re^d$ $(d \ll n)$ by a linear transformation $U \in \Re^{n \times d}$, i.e., $\vec{y}_i = U^T \vec{x}_i$, where principal component analysis (PCA) [20] and linear discriminant analysis (LDA) [32] are two of the most representative methods. PCA projects the data along the direction with the largest variance; whereas LDA finds the projection direction that maximizes the between-class scatter matrix while minimizing the within-class scatter matrix. Manifold learning algorithms aim to explore the local geometrical structure in the low-dimensional manifold embedded in the high-dimensional space, where the representative ones are locally linear embedding (LLE) [35], isometric feature mapping (Isomap) [43], Laplacian Eigenmap (LE) [4], just to name a few. LLE assumes that data points close in the high-dimensional space should also be close in the embedded low-dimensional space and utilizes linear coefficients preserving the local geometry in the high-dimensional space to reconstruct each data point from its neighbors. Similar as LLE, Isomap preserves the intrinsic geometry of data by computing the geodesic distance (shortest path) between pairs of data points. LE was proposed based on the correspondence between the graph Laplacian and the Laplace Beltrami operator on the manifold.

Groeneweg et al. [15] implemented a fast offline building recognition method based on intensity-based region detection [46] and PCA [20]. The algorithm was first tested on ZuBuD [37] by the following steps:

(1) downsample every image in ZuBuD;
(2) detect invariant regions based on local intensity extrema;
(3) fit a parallelogram to each detected region to capture the transformations that affect its appearance;
(4) double the size of the fitted parallelogram to make the region more distinctive;
(5) transform the image contents in each resized region to a fixed size of 10×10;
(6) compute the RGB color values of the pixels in each patch to characterize the region and normalize it by dividing each value by the sum of the intensities of all pixels in the region to make the representation invariant to illumination changes;
(7) apply PCA for compact representation by keeping the first 30 components;
(8) implement linkage clustering to group the features for all images of a building into clusters according to the maximal distance between the instances and characterize each cluster by its centroid. This step can remove repeated regions caused by repetitions (i.e., a row of identical windows) or many views of a building;
(9) build a 100-bin histogram for the r channel and g channel for every building image in the database;
(10) normalize each histogram and store it in the database; and
(11) calculate the chi-square distance between the normalized RGB histogram of the query image and every histogram previously stored in the database. Then the best match is determined by voting each region found within the query image for the building by a weighted majority voting scheme.

This approach reduces the computational cost and storage capacity in a mobile phone platform, but it is sensitive to illumination changes and not invariant to rotation. Considering that ZuBuD is a relatively simple dataset, the Roeterseiland database was constructed by taking pictures of the Roeterseiland complex of the University of Amsterdam to allow the evaluation of the normalized RGB histograms under harder conditions. It consists of images of 7 buildings and the resolution of each image is 160 × 120. The experimental results on this dataset verified that the global color distributions are not discriminative enough in complex environments. Nevertheless, the number of buildings in this dataset is still small, which cannot really serve as a benchmark dataset for investigating various building recognition techniques.

Li and Allinson [26] integrated biologically-inspired feature extraction [40] with dimensionality reduction to construct a biologically-plausible building recognition (BPBR) scheme. Firstly, biologically-inspired features are extracted using a saliency model and a gist model. The saliency model is constructed by extracting visual features at multi-scales and creating a set of feature maps for each image, and then the gist model is constructed by dividing each feature map into a number of sub-regions and describing each map by a gist feature. Afterwards, LDA is used to reduce the dimensionality of the feature vectors and the nearest neighbor rule [11] is applied for classification. Experiments undertaken on SBID demonstrate that this scheme achieves satisfactory results. Extracted features are biologically related to human visual perception and invariant to geometric and photometric transformations – especially robust to different lighting conditions. Moreover, each stage of the scheme requires low computational cost. Based on this scheme, the authors further proposed a relevance feedback (RF)-based building recognition (RFBR) scheme [25] which performs a support vector machine (SVM)-based RF after dimensionality reduction by LDA. The flowchart of the RFBR scheme is given in Fig. 10. It was the first time to embed human–computer interaction in building recognition, resulting in enhanced recognition performance. However, both schemes were evaluated on SBID with relatively small-size images, while their effectiveness cannot be guaranteed for high-definition images.

3.2.2. Clustering-based algorithms

In clustering-based building recognition algorithms, clustering is usually conducted on extracted features (either global or local) using *k*-means or hierarchical clustering which groups the features into different clusters. By exploiting the

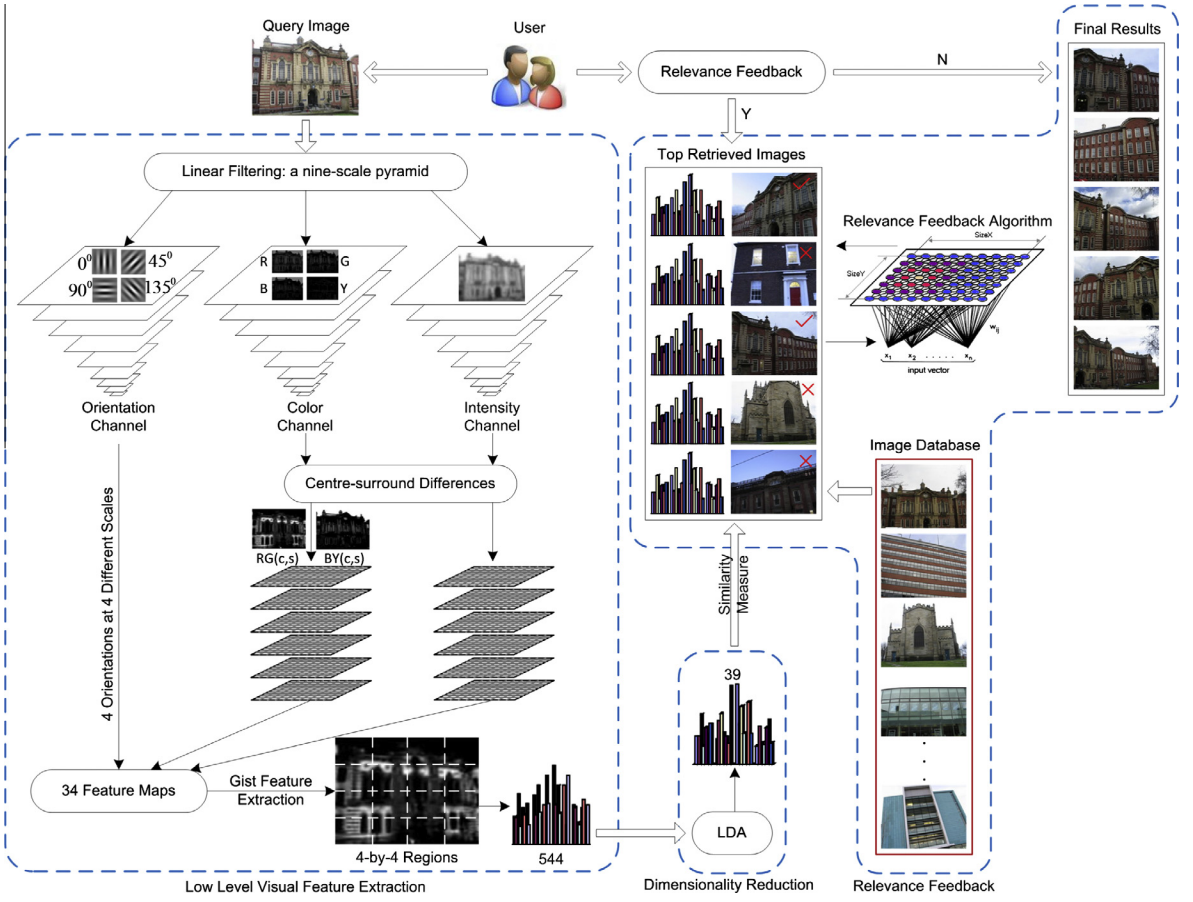

Fig. 10. The flowchart of the RFBR scheme. The figure is taken from [25].

intra-structure of each cluster or modeling the interrelationships between two clusters, the efficiency of building recognition improves.

According to the principles of perceptual grouping [19], a bottom-up processing approach was proposed to explore the semantic interrelationships among different types of primitive image features (e.g., significant edges, junctions and parallel lines). These low-level visual features were hierarchically grouped into intermediate-level structures (e.g., corners, windows, doors, and boundaries of the building) for content-based image retrieval of buildings. In more detail, the following features were hierarchically extracted from an image: (i) straight line segments; (ii) longer linear lines; (iii) "L" junctions; (iv) "U" junctions; (v) parallel lines; (vi) parallel groups; and (vii) significant parallel groups. The obtained parallel lines are grouped into clusters with similar orientations; parallel groups are obtained by grouping overlapping parallel lines which are determined by orthogonal projection; significant parallel groups are extracted based on certain criteria. Finally, a feature vector is described by $X = (x_1, x_2, x_3)^t$ with $i \in \{1, 2, 3\}$ and $x_i \in [0, 1]$, where x_1 denotes the percentage of lines in "L" junctions in the total number of retained lines, x_2 is the percentage of lines in "U" junctions in the total number of retained lines, and x_3 stands for the percentage of lines in significant parallel groups in the total number of retained lines.

Finally, a Bayesian framework is applied to analyze these features and determine the presence of a building in an image. The system was tested on 150 images with a resolution of 640×480 and both the recall and the precision for the building classes are over 80%.

After detecting edges by a Canny edge detector [9] and segmenting them into straight lines using the Object Recognition Toolkit [12], color, orientation, and spatial features of each line segment are integrated and grouped into consistent line clusters [28], i.e., a type of mid-level features, where intra-cluster and inter-cluster relationships are utilized to recognize and locate different buildings in an image. These structural local features are obtained through the construction of color-consistent line clusters, orientation-consistent line clusters, and spatially-consistent line clusters step by step. Firstly, each pixel of an image is classified as one of several dominant colors and every line segment is grouped into one of the color-consistent line clusters based on its color pair, i.e., a dominant color from its left region and the other from its right region. Afterwards, roughly orientation-consistent line clusters are achieved by further classifying every color-consistent line cluster according to the line segments' orientations in the image, where parallel segments of the same orientation are assigned to one orientation-consistent line cluster via finding the peaks in the orientation histogram. In order to rule out the segments from different physical entities, spatial clustering is performed using both vertical and horizontal position histograms to project the line segments to create vertical position clusters and horizontal position clusters, respectively. Consistent line clusters enable both keyword indexing and spatial relationship queries. Nevertheless, the detection rate decreases significantly if the building in an image occupies only a small portion.

Zhang and Košecká [57] proposed a hierarchical building recognition (HBR) 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. Afterwards, an image pixel with its gradient magnitude above a previously defined threshold is assigned to one of the groups (namely left, right, and vertical) if the difference between its gradient direction and the principal vanishing directions is less than some threshold, and then localized color histograms are only computed on these pixels as indexing vectors. Finally, the histograms are matched by the chi-square distance, and the recognition results on ZuBuD [37] are refined by extracting SIFT features and applying a simple probabilistic model to integrate the evidence from individual matches. Because of the fast indexing step using localized color histograms, this method achieves some improvement in efficiency and has attracted the most attention. Nevertheless, it also has limitations: (1) the authors assume that there is only one building in each image to be recognized, which is not always true in real-world cases; (2) although it conducts a fast indexing step, the processing time for extracting many features from color images is still long; (3) its recognition performance is good only when the building is large enough and with simple backgrounds. In consequence, the algorithm is inappropriate for navigation systems that require real-time processing.

Considering visual words as a spatial configuration, a ranking scheme [34] was proposed for searching building facades in a large corpus. Its procedure is given as follows. First, affine-invariant Hessian regions [33] are extracted and described by 128-dimensional SIFT descriptors [29]. These descriptors are then quantized or clustered into a visual vocabulary by approximate *k*-means which employs a forest of 8 randomized kd-trees built over the cluster centers at the beginning of each iteration. In this way, each affine region is mapped to the closest visual word and an image is represented as a bag of visual words. Afterwards, the search engine uses a vector-space model [3] of visual word occurrences in an image and calculates the similarity between the query vector and each image vector in the database according to tf–idf weighting. Finally, the top retrieved results are re-ranked by first estimating a transformation between the query region and each target image and then re-ranking target images based on the discriminability of the spatially verified visual words. The performance of the ranking scheme has been demonstrated on OBD [34]. However, the spatial matching process of the ranking stage adds to the computational burden.

Based on the visual characteristics of landmarks, Zheng et al. [58] proposed a clustering-based landmark recognition method to organize and index landmarks from two sources: (i) GPS-tagged photos in photo sharing websites with their text tags and (ii) travel guide articles from websites. Landmarks mined from these two sources have small overlaps but complement each other: a large number of the geographically calibrated images are visually similar and the names of landmarks can be mined from their corresponding geographic text tags; the landmark list mining can be regarded as a task of text-based named entity extraction from the travel guide corpus. The procedure of the recognition system is given in Fig. 11.

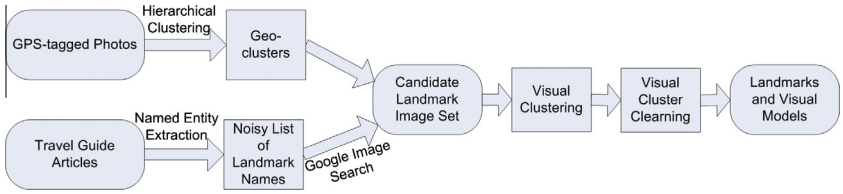

Fig. 11. Clustering-based landmark recognition.

For GPS-tagged photos, the following steps are conducted: (1) perform hierarchical clustering on photos' GPS coordinates to obtain dense geo-clusters; (2) conduct visual clustering on the noisy image set I_1 of each geo-cluster, where I_1 contains photos of one or several adjacent landmarks; (3) extract text tags of each photo in a visual cluster by filtering out stop words and phrases; and (4) compute the frequency of n-grams of all text tags in each visual cluster and the n-grams with the highest frequency are regarded as the landmark name for the visual cluster.

For travel guide articles, the system executes as follows: (1) extract a noisy list of landmark candidates by performing named entity extraction based on the semantic clues embedded in the document structure. That is, the text is classified to be either landmark or non-landmark according to a set of heuristic rules; (2) generate the candidate image set using the mined landmark name associated with each landmark candidate as a query for Google image search; and (3) perform visual clustering on I_2 to exploit true landmark images.

Given candidate image sets $I = I_1 \cap I_2$ which contain potential landmark images from geo-clusters and Google image search, clustering is performed on I to learn true landmark images by analyzing the visual similarity distribution among images. After training an AdaBoost-based photographic vs. non-photographic image classifier and adopting a multi-view face detector [49], visual cluster outlines (non-photographic images) are removed and photos with overly large area of human faces are filtered out, respectively. Afterwards, interest points are detected by Laplacian-of-Gaussian filters [33] and each local region is described by a 118-dimensional Gabor wavelet texture feature. For efficiency, PCA [20] is applied to reduce the feature dimensionality from 118 to 40. Consequently, object matching on all images in the set is performed by comparing the local features for a pair of images and then an undirected weighted region graph is obtained, in which the vertices are matched regions and the edge weight is quantified by its length. Finally, the hierarchical agglomerative clustering [7] is conducted on the graph to discover regions of the same or similar landmarks and graph matching is carried out by utilizing the single linkage inter-cluster distance to define the distance between two regions in order to achieve efficiency. In this work, efficiency is accomplished not only by hierarchical clustering, but also by parallel computing of landmark models on multiple machines and efficient image matching by kd-tree indexing [6] for local features.

Chung et al. [10] utilized sketch-based representations to find major structural components of a building, e.g., windows and doors, for office-building recognition. The scheme detects multi-scale maximal stable extremal regions (MSERs) [30] and describes the normalized MSER patches using histogram of oriented gradients [11]. Afterwards, k -means clustering is applied to group 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. This method was specially designed for building recognition with large viewpoint changes. However, it only focuses on office-building recognition while its performance for other types of buildings has not been demonstrated.

3.2.3. Others

In [2], a rapid window detection and localization method for buildings was introduced for mobile vision systems, where window detection, integrating line grouping, pattern detection, and gradient setting, is considered as a pattern recognition task, respectively. It is based on the extraction of multi-scale Harr-like features followed by a learning stage using a classifier cascade through AdaBoost. The advantage of the proposed method is: instead of detecting every window of a building, only a

Table 2
Recognition rates reported by efficiency approaches on different databases.

Approach	Database	Performance (%)
Groeneweg et al. [15]	ZuBuD	91.0
BPBR [26]	SBID	85.3
RFBR [25]	SBID	93.0
Perceptual grouping [19]	150 Building images	83.7
Consistent line clusters [28]	977 Color images	94.2
Hierarchical building recognition [57]	ZuBuD	95.0
Ranking scheme [34]	OBD (5 M)	95.3
Zheng et al. [58]	5312 Landmarks from 1259 cities	80.8
Sketch-based representations [10]	ZuBuD	81.0
Semantic indexing [2]	ZuBuD	57 ± 19

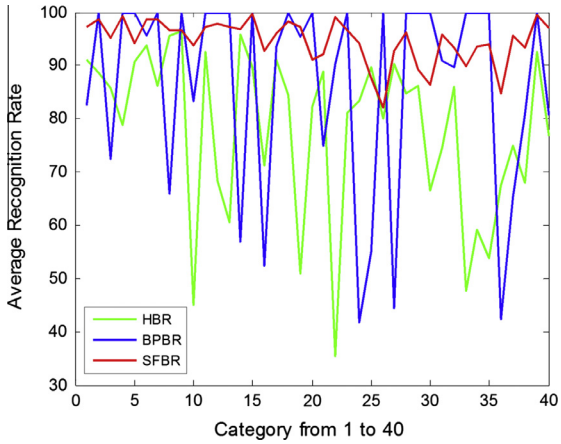

Fig. 12. Comparison results of HBR, BPBR, and SFBR for each category in the Sheffield Building Image Dataset [26]. This figure is taken from [24].

fraction of discriminative windows need to be detected, enabling fast indexing of buildings from mobile imagery in the urban environment.

3.2.4. Summary

The efficiency of building recognition is achieved using either dimensionality reduction or clustering after feature extraction. Efficiency approaches are relatively fast but sometimes cannot achieve satisfactory accuracy [2]. Individual recognition rates of various efficiency approaches are provided in Table 2.

4. Comparison results of representative approaches

Although it is difficult to compare all the above-mentioned building recognition systems systematically, we provide the comparison results of some representative works [24], i.e., hierarchical building recognition (HBR) system [57], biologically-plausible building recognition scheme [26], and steerable filter-based building recognition (SFBR) model [24]. These techniques were conducted on the Sheffield Building Image Dataset (SBID) [26], which is one of the most popularly used datasets for building recognition.

The recognition performance of HBR, BPBR, and SFBR for different individual categories is shown in Fig. 12. As we can see, for most categories in SBID, the performance of the SFBR model is the most stable and better than the others. HBR works well for Category 8 and Category 9. However, its performance is poor for other categories, e.g., Category 10 and Category 22. This indicates that 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, bricks, etc. In SFBR, we use second-order steerable filters [13], 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 why SFBR performs 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.

5. Conclusions and directions for future research

Recent advances in computer vision have promoted new and enhanced techniques for building recognition. As we can see from the selection of work reviewed above, building recognition can be potentially utilized in various computer vision applications, e.g., automatic target detection in surveillance, real-time robot localization and visual navigation, architectural design, and 3D city reconstruction. In this survey, we have reviewed the state-of-the-art techniques of building recognition in urban environments from different aspects: effectiveness and efficiency. Since not all the approaches were tested on the same building image dataset, it is not possible to summarize which approach performs best only through their recognition rates. Nevertheless, we can make the following tentative conclusions for building recognition algorithms.

Feature extraction and classifier design are two key issues in the task of building recognition, where both global features and local features are essential to represent a building since they are two complementary types of features. As a result, most current building recognition systems adopt both of them. Some techniques aim to achieve effectiveness (high recognition

accuracy) but they embed complex building recognition processes which require more computational time; on the other hand, efficiency approaches are relatively fast but sometimes cannot achieve satisfactory accuracy. How can we usefully integrate these approaches? One straightforward way is to integrate the components (i.e., detectors, descriptors, and classifiers) in various algorithms in a sequential order. However, this is not efficient enough, because different algorithms may share more or less the same components. If we explore these algorithms separately, certain algorithmic modules may extract typically redundant information. Therefore, a crude fusion may not improve accuracy but will likely slow down the system. In order to obtain a more efficient system, the investigation for intelligent information fusion or interactive fusion is highly desirable and it is necessary to design improved feature representation and classification methods that are multi-functional.

This survey is extensive, but it does not claim to be complete. There are many other aspects that have not been explored in depth and we list below several promising directions for further research:

(1) A comprehensive benchmark database for performance evaluation needs to be constructed. Although most building recognition techniques were evaluated on one of the following building image datasets, namely ZuBuD, SBID, and OBD, each dataset has its own limitations (please refer to Section 2) and thus are not powerful to serve as a benchmark database for future study. One direction is: except for city building images, other types of data (e.g., live videos, Google Street View images, and satellite images) can be collected and integrated into the existing building recognition datasets. In this way, current datasets can be enlarged and thus become more challenging for real-world applications.
(2) As feature representation can be deemed as one of the most important stages for building recognition, more sophisticated and effective features can be designed to meet the requirement of different applications. Moreover, instead of using some specific local features (e.g., SIFT [29]) to detect interest points and describe local regions, machine-learned features, which are more adaptive for this particular task, can be introduced into the building recognition scheme.
(3) The curse of dimensionality is a significant problem in recognition tasks, where subspace learning-based dimensionality reduction methods play a dominant role to alleviate the problem. However, rather than using classical linear subspace methods, e.g., PCA or LDA, more recent and advanced dimensionality reduction techniques can be adopted according to the preference to specific applications.
(4) As building recognition is usually regarded as a content-based image retrieval problem, more advanced relevance feedback techniques can be adopted to bridge the gap between low-level visual features and high-level image concepts through preprocessing on accumulated user log files or statistical analysis of the log files, in order to obtain significant results to accelerate the retrieval process as well as improving the retrieval effectiveness.
(5) Nowadays, much attention has been paid to 3D reconstruction. Building recognition schemes can be extended to the 3D modeling in city planning and design for synthesizing scenes in computer games.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S.M. Seitz, R. Szeliski, Reconstructing Rome, IEEE Comput. 43 (6) (2010) 40–47.
[2] H. Ali, G. Paar, L. Paletta, Semantic indexing for visual recognition of buildings, in: Proc. Int'l. Symposium on Mobile Mapping Technology, 2007, pp. 28–31.
[3] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, 1999. ISBN: 020139829.
[4] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inform. Process. Syst. 14 (2002) 585–591.
[5] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, NJ, 1961.
[6] J.L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM 18 (9) (1975) 509–517.
[7] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[8] Y-L. Boureau, J. Ponce, Y. Lecun, A theoretical analysis of feature pooling in visual recognition, in: Int'l. Conf. Machine Learning, 2010.
[9] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (1986) 679–698.
[10] Y-C. Chung, T.X. Han, Z. He, Building recognition using sketch-based representations and spectral graph matching, in: IEEE Int'l. Conf. Computer Vision, 2009.
[11] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.
[12] A. Etamadi, Robust segmentation of edge data, in: Int'l. Conf. Image Processing and its Applications, 1992, pp. 311–314.
[13] W. Freeman, E. Adelson, The design and use of steerable filters, IEEE Trans. Pattern Anal. Mach. Intell. 13 (9) (1991) 891–906.
[14] T. Goedeme, T. Tuytelaars, L.V. Gool, Fast wide baseline matching for visual navigation, in: IEEE Conf. Computer Vision and Pattern Recognition, 2004, pp. 24–29.
[15] N. Groeneweg, B. Groot, A. Halma, B. Quiroga, M. Tromp, F. Groen, A fast offline building recognition application on a mobile telephone, in: Advanced Concepts for Intelligent Vision Systems, 2006, pp. 1122–1132.
[16] S. Gunal, R. Edizkan, Subspace based feature selection for pattern recognition, Inform. Sci. 178 (19) (2008) 3716–3726.
[17] C. Harris, M. Stephens, A combined corner and edge detector, in: Alvey Vision Conf., 1988, pp. 147–151.
[18] R. Hutchings, W. Mayol-Cuevas, Building Recognition for Mobile Devices: Incorporating Positional Information with Visual Features, CSTR-06-017, Computer Science, University of Bristol, 2005.
[19] Q. Iqbal, J.K. Aggarwall, Applying perceptual grouping to content-based image retrieval: building images, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1999, pp. 42–48.
[20] I.T. Jolliffe, Principal Component Analysis, second ed., Springer, 2002.
[21] S. Jones, L. Shao, Content-based retrieval of human actions from realistic video databases, Inform. Sci. 236 (2013) 56–65.
[22] Y. Kim, K. Lee, K. Choi, S.I. Cho, Building recognition for augmented reality based navigation system, in: IEEE Int'l. Conf. Computer and Information Technology, 2006.
[23] J. Li, N.M. Allinson, A comprehensive review of current local features for computer vision, Neurocomputing 71 (10–12) (2008) 1771–1787.
[24] J. Li, N.M. Allinson, Building recognition using local oriented features, IEEE Trans. Ind. Inform. 9 (3) (2013) 1697–1704.
[25] J. Li, N.M. Allinson, Relevance feedback-based building recognition, in: SPIE Visual Communications and Image Processing, vol. 7744, 2010.
[26] J. Li, N.M. Allinson, Subspace learning-based dimensionality reduction in building recognition, Neurocomputing 73 (1–3) (2009) 324–330.
[27] J. Li, N. Allinson, D. Tao, X. Li, Multitraining support vector machine for image retrieval, IEEE Trans. Image Process. 15 (11) (2006) 3597–3601.

[28] Y. Li, L.G. Shapiro, Consistent line clusters for building recognition in CBIR, in: Proc. IEEE Int'l. Conf. Pattern Recognition, 2002, pp. 952–956.
[29] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[30] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: British Machine Vision Conf., 2002.
[31] H. Mayer, Automatic object extraction from aerial imagery – a survey focusing on buildings, Comput. Vision Image Understanding 74 (2) (1999) 138–149.
[32] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley-Interscience, New York, 1992.
[33] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, Int. J. Comput. Vision 1 (60) (2004) 63–86.
[34] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: IEEE Conf. Computer Vision and Pattern Recognition, 2007.
[35] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[36] K. Schindler, J. Bauer, A model-based method for building reconstruction, in: IEEE Int'l. Workshop on Higher-Level Knowledge in 3D Modeling and Motion, Analysis, 2003, pp. 74–82.
[37] T.S.H. Shao, L.V. Gool, Zubud-Zurich buildings database for image based recognition, Technique Report No. 260, Swiss Federal Institute of Technology, 2003.
[38] H. Shao, T. Svoboda, T. Tuytelaars, L.J.V. Gool, HPAT indexing for fast object/scene recognition based on local appearance, in: Int'l. Conf. Image and Video Retrieval, 2003, pp. 71–80.
[39] L. Shao, T. Kadir, M. Brady, Geometric and photometric invariant distinctive regions detection, Inform. Sci. 177 (4) (2007) 1088–1122.
[40] C. Siagian, L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2) (2007) 300–312.
[41] N. Snavely, Steven M. Seitz, Richard Szeliski, Photo tourism: exploring photo collections in 3D, ACM Trans. Graph. (2006) 835–846.
[42] W. Suleiman, T. Joliveau, E. Favier, Buildings Recognition and Camera Localization using Image Texture Description, 2011.
[43] J. Tenenbaum, V. Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[44] A. Treisman, G. Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (1980) 97–137.
[45] H. Trinh, D.-N. Kim, K.-H. Jo, Facet-based multiple building analysis for robot intelligence, J. Appl. Math. Comput. 205 (2) (2008) 537–549.
[46] T. Tuytelaars, Van Gool, Wide baseline stereo based on local affinely invariant regions, in: British Machine Vision Conf., 2000.
[47] M.M. Ullah, A. Pronobis, B. Caputo, J. Luo, R. Jensfelt, H.I. Christensen, Towards robust place recognition for robot localization, in: IEEE Int'l. Conf. Robotics and Automation, 2008, pp. 530–537.
[48] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1995.
[49] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. of Conf. on Computer Vision and Pattern Recognition, 2001, pp. 511–518.
[50] M. Xu, M. Petrou, M. Jahangiri, Component identification in the 3D model of a building, in: Int'l. Conf. Pattern Recognition, 2010, pp. 3061–3064.
[51] Z-J. Zha, X-S. Hua, T. Mei, J. Wang, G-J. Qi, Z. Wang, Joint multi-label multi-instance learning for image classification, in: IEEE Conf. Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[52] Z-J. Zha, M. Wang, Y-T. Zheng, Y. Yang, R. Hong, T-S. Chua, Interactive video indexing with statistical active learning, IEEE Trans. Multimedia 14 (1) (2012) 17–27.
[53] Z-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, Visual query suggestion, in: ACM Int'l. Conf. Multimedia, 2009, pp. 15–24.
[54] Z-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T-S. Chua, X-S. Hua, Visual query suggestion: towards capturing user intent in internet image search, ACM Trans. Multimedia Comput. Commun. Appl. 6 (3) (2010).
[55] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, Q. Tian, Discovering discriminative graphlets for aerial image categories recognition, IEEE Trans. Image Process. 22 (12) (2013) 5071–5084.
[56] L. Zhang, M. Song, X. Liu, L. Sun, C. Chen, J. Bu, Recognizing architecture styles by hierarchical sparse coding of blocklets, Inform. Sci. 254 (2014) 141–154.
[57] W. Zhang, J. Košecká, Hierarchical building recognition, Image Vision Comput. 25 (5) (2007) 704–716.
[58] Y-T. Zheng, M. Zhao, Y. Song, H. Adam, Tour the world: building a web-scale landmark recognition engine, in: IEEE Conf. Computer Vision and Pattern Recognition, 2009, pp. 1085–1092.
[59] robotics.youngester.com.
[60] http://web.eee.sztaki.hu/home4/.
[61] apnagharit.wordpress.com.
[62] www.stanford.edu.
[63] http://www.flickr.com/.