Neurocomputing 73 (2009) 324–330

Contents lists available at ScienceDirect

Neurocomputing

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

Subspace learning-based dimensionality reduction in building recognition

Jing Li *, Nigel M. Allinson

Vision and Information Engineering Research Group Department of Electronic and Electrical Engineering Mappin Street, University of Sheffield, Sheffield, S1 3JD, UK

ARTICLE INFO

Article history:
Received 9 May 2009
Received in revised form
12 August 2009
Accepted 16 August 2009
Communicated by X. Li
Available online 4 October 2009

Keywords:
Subspace learning
Building recognition
Biologically-inspired feature extraction
Gist features
Dimensionality reduction

ABSTRACT

Building recognition is a relatively specific recognition task in object recognition, which is challenging since it encounters rotation, scaling, illumination changes, occlusion, etc. Subspace learning, which dominates dimensionality reduction, has been widely exploited in computer vision research in recent years. It consists of classical linear dimensionality reduction methods, manifold learning, etc. To explore how different subspace learning algorithms affect building recognition, some representative algorithms, i.e., principal component analysis, linear discriminant analysis, locality preserving projections (unsupervised/supervised), and semi-supervised discriminant analysis, are applied for dimensionality reduction. Moreover, a building recognition scheme based on biologically-inspired feature extraction is proposed in this paper. Experiments undertaken on our own building database demonstrate that the proposed scheme embedded with subspace learning can achieve satisfactory results.

1. Introduction

Object recognition is reaching a certain maturity in computer vision research, where a number of algorithms [9–12,23–25,28,31,48] have been proposed that are based on either global or local features. However, most algorithms were designed for categorization, e.g., scene or diverse objects [19,21,34,44,47,49], but the classification within a specific class has been somewhat neglected. Building recognition is such an intra-class classification task that aims to distinguish different buildings in a large-scale image database.

Building recognition can be utilised in many practical application areas, such as architectural design, building labelling in videos, robot vision or localization [47], and mobile device navigation [1,17]. However, this task is more challenging compared to general object recognition since building images may be taken from different viewpoints, under different lighting conditions, or suffer from occlusion from trees, moving vehicles, other buildings or themselves. As a result, little attention has been paid to this specific recognition task. In [18], perceptual grouping in a hierarchical way was applied to explore the semantic relationships among low-level visual features for content-based image retrieval (CBIR) for buildings. Consistent line clusters, a type of mid-level feature, were put forward in [27] for building recognition in CBIR. After extracting colour, orientation, and spatial information for each line segment, they were grouped into different consistent clusters, and the intra-cluster and inter-cluster relationships can be utilised to recognise different buildings. Zhang and Kosecká [55] proposed a building recognition system based on vanishing point detection and localized colour histograms. Because of the fast indexing step using localized colour histograms, this method achieved some improvement in efficiency. Hutchings and Mayol-Cuevas [17] applied the Harris corner detector [13] to extract interest points for matching buildings in the world space for mobile device.

Nevertheless, the building recognition systems mentioned above suffer from some limitations, namely (i) they are based on the detection of low-level visual features, e.g., line segments, vanishing points, etc. However, low-level features cannot reveal the truly semantic concepts of images, which limit the representational performance; and (ii) recognition is conducted on pairs of raw high-dimensional feature vectors, resulting in high computational cost and memory requirements.

To this end, we propose a building recognition scheme to address these problems. The scheme consists of the three stages: (i) feature representation; (ii) dimensionality reduction; and (iii) classification. Features that are biologically present in the human visual perception are extracted using *saliency* and *gist* models [34], where the saliency model is constructed by extracting visual information from a raw image and the gist model is based on the extracted visual features. In detail, visual features are extracted at multi-scales and a set of feature maps is created for each image, resulting in the saliency model. After that, the gist model is constructed by dividing each feature map into a number of sub-regions and describing each map by a gist feature.

* Corresponding author.
E-mail addresses: elq06jl@sheffield.ac.uk (J. Li), n.allinson@sheffield.ac.uk (N.M. Allinson).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.08.016

In order to reduce the computational cost while preserving most of the discriminative information for recognition, we reduce the original higher dimensional feature vectors to a much lower dimensional feature space by dimensionality reduction. This is a significant step in computer vision applications since extracted features of visual objects are always of high dimension, from several hundreds to tens of thousands, which may contain some redundant and useless information. If they are directly utilised for subsequent operations, e.g., classification, not only does it require high computational cost and large memory, but also it encounters the curse of dimensionality [3], which degrades the final performance when the dimension exceeds a certain value. Fortunately, dimensionality reduction can make data more compact and is able to alleviate the problems mentioned above. Wherein, subspace learning-based dimensionality reduction, which finds a projection that reduces the high-dimensional feature space to a lower dimensional subspace, is getting more and more popularity in data mining and computer vision applications. A number of subspace learning methods, widely applied in biometrics [37,50–54] and multimedia information retrieval [14,15,38–42], have been proposed in the literature. Among them, conventional linear dimensionality reduction techniques [20,30] and manifold learning algorithms [2,5,16,33,43] dominate.

Principal component analysis (PCA) and linear discriminant analysis (LDA) are the most significant linear dimensionality reduction methods. PCA projects data along the direction of the largest variance; while LDA maximizes the ratio of the between-class scatter matrix to the within-class scatter matrix. However, PCA is unsupervised since it does not consider any label information, whereas LDA is a supervised learning method that takes the class label information into account.

Manifold learning algorithms aim at exploring the intrinsic geometries in the manifold embedded in high-dimensional space. Locally linear embedding (LLE) [33] is a nonlinear learning algorithm, which assumes data points close in the high-dimensional space should also be close in the embedded lower dimensional space. Therefore, the linear coefficients, preserving the local geometry in the high-dimensional space, are utilised to reconstruct each data point from its nearest neighbours. ISOMAP [43] finds the shortest geodesic distance between pairs of data points. Laplacian eigenmaps (LE) [2] constructs an adjacency map with weighted edges from nearest neighbours. Locality preserving projections (LPP) [5,16] applies the main idea in LE [2], but it is the linearization approximation of LE. Semi-supervised discriminant analysis (SDA) [4] explores the manifold structure of data points by considering both labelled examples and unlabelled examples.

To evaluate how different subspace learning-based dimensionality reduction methods perform for the building recognition task, some representative algorithms, i.e., the principal component analysis (PCA) [20], locality preserving projections (LPP) [16], supervised LPP (SLPP) [5], linear discriminant analysis (LDA) [30], and semi-supervised discriminant analysis (SDA) are tested on our own building database, which incorporates various challenges including scaling, rotation, illumination changes, viewpoint changes, and camera-shake. After dimensionality reduction, building recognition is implemented by the nearest neighbour rule [6].

The advantages of our system are: (i) the extracted features are biologically related to the human visual perception; (ii) features are invariant to scaling, rotation, illumination changes, and occlusion, and especially robust to different lighting conditions; and (iii) each stage of our system requires low computational cost.

The organization of this paper is that we describe the biological-based features representation in Section 2; representative subspace learning-based dimensionality reduction algorithms are introduced in Section 3; recognition performance on our own database is detailed in Section 4; while Section 5 concludes.

2. Feature representation

A large body of psychological research [45] supports the view that humans are able to grasp the holistic information in an image – also known as the gist [32] of a scene – (e.g., indoor or outdoor, approximate locations, dominative colours) by glancing at it for just a few seconds. Inspired by the biological idea in [34], we utilise gist features as well as low-level visual features for building recognition that are known to provide more discriminative information for this type of recognition task. The key idea of feature representation is to first construct saliency feature maps, followed by generating a gist feature for each of them.

Saliency feature maps are constructed based on low-level visual features, including both global and local features, which are extracted in parallel. Global features are obtained by extracting intensity and colour information at different scales; local features are acquired by generating Gabor features [7,8] at several different scales and orientations. Although the features mentioned above have been previously utilised for other recognition tasks [35,36], it is the first time they are combined together for building recognition. After low-level feature extraction, a gist feature is extracted for each saliency map. To integrate low-level visual features with gist features, three major steps are conducted: (i) linear filtering; (ii) visual feature extraction; and (iii) gist feature generation. Each step will be detailed in the following sections, while the entire recognition scheme is represented in Fig. 1.

2.1. Linear filtering

Each input image $I(\vec{p})$ with $\vec{p}=[x,y]^T$, is linearly filtered to give a Gaussian pyramid of nine scales $I(\vec{p};\sigma)$ (σ=0,…,8). Afterwards, the centre-surround operation [19], widely used in modelling the receptive fields in the human visual system, is conducted by computing pixel differences across scales

$$I_{c,s}(\vec{p}) = |I(\vec{p};c) - I(\vec{p};s)|, \tag{1}$$

where $I(\vec{p};c)$ denotes a pixel at a centre scale c=2,3,4, $I(\vec{p};s)$ is its corresponding pixel at a surround scale $s=c+d$ with d=3,4, and $I(\vec{p};s)$ is interpolated to the centre scale so that pixel difference maps are obtained.

For simplicity, the centre-surround operation is reformulated as

$$I_{c,s} = |I_c \Theta I_s|, \tag{2}$$

It should be noted that linear filtering is a prerequisite step for all visual feature extractions, but the centre-surround is only performed for intensity and colour information extraction, which creates on centre-surround difference maps of $I_{2,5}$, $I_{2,6}$, $I_{3,6}$, $I_{3,7}$, $I_{4,7}$, and $I_{4,8}$.

2.2. Visual feature extraction

Generally, there are many robust types of features for object recognition, such as colour, shape, texture, intensity, motion, etc. Applying more types of features may permit greater accuracy for classification; on the other hand, it means more computational complexity is required. To balance the accuracy and computational cost, only three types of visual features are adopted for the building recognition task, i.e., intensity, colour, and orientation, respectively. This results in three main visual channels: intensity channel, colour channel, and orientation channel. To simplify the