Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1021–1031
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Robust automatic selection of optimal views in multi-view free-form object recognition

Farzin Mokhtarian*, Sadegh Abbasi

Centre for Vision Speech and Signal Processing, Department of Electronic & Electrical Engineering, University of Surrey, Guildford, Surrey GU2 7XH, UK

Received 25 June 2003; accepted 15 November 2004

Abstract

This paper addresses the issue of automatic selection of the best and the optimum number of representative views for each object in a database that can enable the accurate recognition of that object from any single arbitrary view of the object. The object boundary in each view is represented by its curvature scale space (CSS) image. The CSS representation has been selected for MPEG-7 standardisation as a contour shape descriptor.

The paper also presents a novel method for fusion of results from combined shape descriptors. The utilisation of this method for multi-view three-dimensional (3-D) object recognition has been explored. The object boundary of each view is represented effectively using the CSS technique, moment invariants and Fourier descriptors. It has been shown that the results obtained from the fusion method are superior to the results obtained from any single technique.

The method has been tested on a collection of free-form 3-D objects. Each object has been modelled using an optimal number of silhouette contours obtained from different viewpoints. This number varies depending on the complexity of the object and the measure of expected accuracy. A comprehensive analysis of the performance of the system has been given.

Keywords: Multi-view object recognition; Automatic view selection; Shape representation; Curvature scale space; Fourier descriptors; Moment invariants

1. Introduction

There are two major methods of data acquisition in three-dimensional (3-D) object recognition. In laser-based systems, range data is produced which essentially records the distances between the camera and different points of the object. In image-based systems, a CCD camera is used to produce two-dimensional (2-D) images from the object. In both methods, the 3-D object is finally represented by a number of features extracted either from the range data or from 2-D images. While range data provides more accurate information about the surface of the 3-D object, laser-based systems are more expensive and the related recognition methods are more time-consuming. The computational cost of a method has recently become much more important in search and retrieval from large databases. Although hierarchical methods, and different indexing techniques have been introduced to narrow down the search space in such applications, the need for rapid matching methods to find the best matches among the remaining candidates still exists.

A large number of 3-D object representation methods have been introduced in the object recognition literature. Note that in this paper we are only interested in representation techniques which are intended for use in object recognition. They can be categorised based on the type of

* Corresponding author. Tel.: +44 1483 686035; fax: +44 1483 686031.
E-mail address: f.mokhtarian@surrey.ac.uk (F. Mokhtarian).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.11.021

descriptors they compute to represent the 3-D object. It should also be mentioned that some methods impose restrictions on the classes of geometric objects that can be handled.

The aspect graph is a viewer-centred representation of a 3-D object that enumerates all the topologically distinct views of that object. Constructing an aspect graph requires partitioning the viewpoint space into view-equivalent cells by a number of visual event surfaces [1–4]. However, construction of an aspect graph representation for a free-form object is in general complex and computationally intensive. Furthermore, topologically equivalent views of a 3-D object may still appear different from each other which would make robust recognition difficult. In [5], parallel lines and ellipses are used to describe different viewpoints of an object. A strategy is suggested to recognise an object from an unknown viewpoint. The method is based on earlier works by Brooks [6] which has been modified later on by others [7,8]. In Ref. [9], a model is established from a large number of viewpoints taken from a video sequence. The input of the system is also a video sequence of an unknown object. The system first builds a 2-D representation of the object. If the representation matches one of the objects of the database, it is modified based on new information extracted from the new sequence. Otherwise the object is recognised as a new one and its representation is stored. The number of views may be reduced with further preprocessing [10]. Representation with multiple views and recognition using a single view was also proposed in Ref. [11]. The number of viewpoints used to represent a complex object was down from about 2000 in Ref. [9] to 20–30 as reported in Ref. [11]. Shape [12,13] and colour [14] features have also been used in 3-D object representation.

Multi-view representations have not yet successfully dealt with the following issues [15]:

- What is the optimal number of views?
- How to select the optimal views?

In this article, we propose a method for automatic selection of optimal views of an object starting from a set of greyscale images of that object corresponding to as many different views of the object as possible. In order to represent an object efficiently, we eliminate similar views and select a relatively small number of views using an optimisation algorithm [16]. The goal of this representation method is subsequent recognition of the objects under consideration. The number of views used to represent each object varies from 5 to 25 depending on the complexity of the object and the measure of expected accuracy. To identify an unknown object from a single viewpoint, its representation is matched with all images of the database and the best matches are retrieved and displayed.

In order to represent each view of the object, we need a contour shape descriptor. Since the camera is allowed to change its viewpoint with respect to the object, the resulting boundary of the object may be deformed. The deformation can be approximated by affine transformation and therefore the descriptor must be affine invariant. A number of shape representations have been proposed to recognise shapes even under affine transformation. Some of them are the extensions of well-known methods such as Fourier descriptors [17] and moment invariants [18,19]. Affine invariant scale space is introduced in Ref. [20]. It generalises the definition of curvature to introduce *affine curvature*. This is a curve evolution method which is proved to have similar properties as curvature evolution [21,22], as well as being affine-invariant. However, an explicit shape representation has yet to be introduced based on the theory of affine-invariant scale space [23]. The prospective shape representation might be computationally complex as the definition of affine curvature involves higher order derivatives. A number of shape representation techniques are based on level-set methods [24,25] and volumetric diffusion [26]. These representations suffer from inefficiency and lack of robustness with respect to occlusion. Other techniques based on curve evolution [27] are more suitable for applications other than shape representations.

In our system, each view of an object is represented by the locations of the maxima of its curvature scale space (CSS) image. The representation has already been used to represent shapes of boundaries in similarity retrieval applications [28] and proved to be robust under general affine transforms [29]. It has also been selected for MPEG-7 standardisation as a contour shape descriptor.

In this paper we also explore the fusion of results from combined shape descriptors for automatic selection of optimal views in multi-view 3-D object recognition. In both view selection and recognition stages, the results of three different descriptors are combined to achieve the best performance. Since each descriptor captures some features of the contour, a combination of well-selected shape descriptors will result in better representation of the silhouette as well as the 3-D object. For example, while CSS descriptor captures local features of the shape, Fourier descriptors and moment invariants convey more information about the global features of a shape. Our optimal view selection method is independent of shape descriptors. In fact, it can also be used to select optimal views based on other features such as colour and texture. Using each shape descriptor, the view selection method returns a set of views. These sets are combined at the next stage to find the final set of optimal views. To identify an unknown object from a single viewpoint, its combined representation is matched with the representations of all objects in the database and the best matches are retrieved and displayed. In our experiments with a collection of 15 toy aircrafts of different shapes, we observed that the results obtained from the fusion method are superior to the results obtained from any single technique.

The following is the organisation of the remainder of this paper. In Section 2, the CSS image is reviewed and CSS matching is briefly explained. Fourier descriptors and moment invariants are also reviewed in this section.