PERGAMON

Pattern Recognition 36 (2003) 2187–2196

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

The choice of vantage objects for image retrieval

Christian Hennig[a,b,*], Longin Jan Latecki[c]

[a] *ETH Zürich, Seminar für Statistik, Zürich CH-8092, Switzerland*
[b] *Fachbereich Mathematik, Universität Hamburg, Hamburg 20146, Germany*
[c] *Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA*

Received 2 November 2001; received in revised form 26 August 2002; accepted 7 October 2002

Abstract

Suppose that we have a matrix of dissimilarities between n images of a database. For a new image, we would like to select the most similar image of our database. Because it may be too expensive to compute the dissimilarities for the new object to all images of our database, we want to find $p \ll n$ "vantage objects" (Pattern Recognition 35 (2002) 69) from our database in order to select a matching image according to the least Euclidean distance between the vector of dissimilarities between the new image and the vantage objects and the corresponding vector for the images of the database. In this paper, we treat the choice of suitable vantage objects. We suggest a loss measure to assess the quality of a set of vantage objects: For every image, we select a matching image from the remaining images of the database by use of the vantage set, and we average the resulting dissimilarities. We compare two classes of choice strategies: The first one is based on a stepwise forward selection of vantage objects to optimize the loss measure. The second is to choose objects as representative as possible for the whole range of the database.

Keywords: Cross-validation; Leave-one-out; Stepwise forward selection; Shape similarity

1. Introduction

In this paper, we deal with the following problem: Suppose that we have a database of n images (objects). The information about the images is given in form of $n(n-1)/2$ dissimilarities between them. For a new image, we would like to select the most similar image from our database. This requires the computation of n dissimilarities. Suppose that there is some computational effort to calculate a single dissimilarity, and that it is feasible to calculate a small number of dissimilarities, say 20 or 40, but not all n. Vleugels and Veltkamp [1] suggest the following strategy: Choose a suitable number p of objects from the database as "vantage objects". Calculate the dissimilarities between the new object and the vantage objects. Interpret every object in the database, as well as the new object, as a p-dimensional vector in the Euclidean space, namely as the vector of dissimilarities to the vantage objects. Select the object in the database, whose vector of dissimilarities to the vantage objects has the smallest Euclidean distance to the vector of the new image. This means that for the classification of the new image only p dissimilarity calculations are needed.

The question arises how to choose the vantage objects. Vleugels and Veltkamp [1] suggest some heuristic strategies. We present here a data driven approach to measure the quality of a set of vantage objects by means of a loss function and we suggest and compare some strategies to find high quality sets. If p would be so small that evaluation of the loss of all $\binom{n}{p}$ vantage sets of size p would be possible, the loss function could be optimized directly.

* Corresponding author. Seminar fur Statistic, ETH Zurich (LEO), Zurich CH-8092, Switzerland. Tel.: +41-632-6184; fax: +41-632-1228.

E-mail addresses: henni@math.uni-hamburg.de (C. Hennig), latecki@temple.edu (L.J. Latecki).

URLs: http://www.math.uni-hamburg.de/home/hennig, http://www.cis.temple.edu/latecki/

0031-3203/03/$30.00
PII: S0031-3203(02)00326-6