Robust Face Recognition for Data Mining

Brian C. Lovell
The University of Queensland, Australia

Shaokang Chen
The University of Queensland, Australia

INTRODUCTION

While the technology for mining text documents in large databases could be said to be relatively mature, the same cannot be said for mining other important data types such as speech, music, images and video. Yet these forms of multimedia data are becoming increasingly prevalent on the Internet and intranets as bandwidth rapidly increases due to continuing advances in computing hardware and consumer demand. An emerging major problem is the lack of accurate and efficient tools to query these multimedia data directly, so we are usually forced to rely on available metadata, such as manual labeling. Currently the most effective way to label data to allow for searching of multimedia archives is for humans to physically review the material. This is already uneconomic or, in an increasing number of application areas, quite impossible because these data are being collected much faster than any group of humans could meaningfully label them — and the pace is accelerating, forming a veritable explosion of non-text data. Some driver applications are emerging from heightened security demands in the 21st century, post-production of digital interactive television, and the recent deployment of a planetary sensor network overlaid on the Internet backbone.

BACKGROUND

Although they say a picture is worth a thousand words, computer scientists know that the ratio of information contained in images compared to text documents is often much greater than this. Providing text labels for image data is problematic because appropriate labeling is very dependent on the typical queries users will wish to perform, and the queries are difficult to anticipate at the time of labeling. For example, a simple image of a red ball would be best labeled as sports equipment, a toy, a red object, a round object, or even a sphere, depending on the nature of the query. Difficulties with text metadata have led to researchers concentrating on techniques from the fields of Pattern Recognition and Computer Vision that work on the image content itself.

A motivating application and development testbed is the emerging experimental planetary scale sensor Web, IrisNet (Gibbons, Karp, Ke, Nath, & Sehan, 2003). IrisNet uses Internet connected desktop PCs and inexpensive, off-the-shelf sensors such as Webcams, microphones, temperature, and motion sensors deployed globally to provide a wide-area sensor network. IrisNet is deployed as a service on PlanetLab (www.planet-lab.org), a worldwide collaborative network environment for prototyping next generation Internet services initiated by Intel Research and Princeton University that has 177 nodes as of August, 2004. Gibbons, Karp, Ke, Nath, & Sehan envisage a worldwide sensor Web in which many users can query, as a single unit, vast quantities of data from thousands or even millions of planetary sensors. IrisNet stores its sensor-derived data in a distributed XML schema, which is well-suited to describing such hierarchical data as it employs self-describing tags. Indeed the robust distributed nature of the database can be most readily compared to the structure of the Internet DNS naming service.

The authors give an example of IrisNet usage where an ecologist wishes to assess the environmental damage after an oil spill by locating beaches where oil has affected the habitat. The query would be directed toward a coastal monitoring service that collects images from video cameras directed at the coastline. The ecologist would then receive images of the contaminated sites as well at their geographic coordinates. Yet the same coastal monitoring service could be used simultaneously to locate the best beaches for surfing. Moreover, via stored trigger queries, the sensor network could automatically notify the appropriate lifeguard in the event of detecting dangerous rips or the presence of sharks.

A valuable prototype application that could be deployed on IrisNet is wide area person recognition and location services. Such services have existed since the emergence of human society to locate specific persons when they are not in immediate view. For example, in a crowded shopping mall, a mother may ask her child, "Have you seen your sister?" If there were a positive response, this may then be followed by a request to know the time and place of the last sighting, or perhaps by a request to go look for her. Here the mother is using the eyes, face recognition ability, memory persistence, and mobility of the child to perform the search. If the search fails, the mother may then ask the mall manager to give a "lost child" announcement over the public address system. Eventually the police may be asked to employ these human search services on a much wider scale by showing a photograph of the missing child on the television to ask the wider community for assistance in the search.

On the IrisNet the mother could simply upload a photograph of her child from the image store in her mobile phone and the system would efficiently look for the child in an ever-widening geographic search space until contact was made. Clearly in the case of IrisNet, there is no possibility of humans being employed to identify all the faces captured by the planetary sensor Web to support the search, so the task must be automated. Such a service raises inevitable privacy concerns, which must be addressed, but the service also has the potential for great public good as in this example of reuniting a worried mother with her lost child.

In addition to person recognition and location services on a planetary sensor Web, another interesting commercial application of face recognition is a system to semi-automatically annotate video streams to provide content for digital interactive television. A similar idea was behind the MIT MediaLab Hypersoap project (Agamanolis & Bove, 1997). In this system, users touch images of objects and people on a television screen to bring up information and advertising material related to the object. For example, a user might select a famous actor and then a page would appear describing the actor, films in which they have appeared, and the viewer might be offered the opportunity to purchase copies of their other films. Automatic face recognition and tracking would greatly simplify the task of labeling the video in post-production — the major cost component of producing such interactive video.

Now we will focus on the crucial technology underpinning such data mining services — automatically recognizing faces in image and video databases.

MAIN THRUST

Robust Face Recognition

Robust face recognition is a challenging goal because of the gross similarity of all human faces compared to large differences between face images of the same person due to variations in lighting conditions, view point, pose, age, health, and facial expression. An ideal face recognition system should recognize new images of a known face and be insensitive to nuisance variations in image acquisition. Yet, differences between images of the same face (intraclass variation) due to these nuisance variations in image capture are often greater than those between different faces (interclass variation) (Adinj, Moses, & Ulman, 1997), making the task extremely challenging. Most systems work well only with images taken under constrained or laboratory conditions where lighting, pose, and camera parameters are strictly controlled. This requirement is much too strict to be useful in many data mining situations when only a few sample images are available, such as in recognizing people from surveillance videos from a planetary sensor Web or searching historic film archives.

Recent research has been focused on diminishing the impact of nuisance factors on face recognition. Two main approaches have been proposed for illumination invariant recognition. The first is to represent images with features that are less sensitive to illumination change (Yilmaz & Gokmen, 2000; Gao & Leung, 2002), such as using the edge maps of an image. These methods suffer from robustness problems because shifts in edge locations resulting from small rotation or location errors significantly degrade recognition performance. Yilmaz and Gokmen (2000) proposed using "hills" for face representation; others use derivatives of the intensity (Edelman, Reisfeld, & Yeshurun, 1994; Belhumeur & Kriegman, 1998). No matter what kind of representation is used, these methods assume that features do not change dramatically with variable lighting conditions. Yet this is patently false as edge features generated from shadows may have a significant impact on recognition.

The second main approach is to construct a low dimensional linear subspace for the images of faces taken under different lighting conditions. This method is based on the assumption that images of a convex Lambertian object under variable illumination form a convex cone in the space of all possible images (Belhumeur & Kriegman, 1998). Once again, it is hard for these systems to deal with cast shadows. Furthermore, such systems need several images of the same face taken under different lighting source directions to construct a model of a given face — in data mining applications it is often impossible to obtain the required number of images. Experiments performed by Adinj, Moses, and Ulman (1997) show that even with the best image representations using illumination insensitive features and the best distance measurement, the misclassification rate is often more than 20%.

As for expression invariant face recognition, this is still an open problem for machine recognition and is

Related Content