Real-Time Face Detection and Classification for ICCTV

Brian C. Lovell
The University of Queensland, Australia

Shaokang Chen
NICTA, Australia

Ting Shan
NICTA, Australia

INTRODUCTION

Data mining is widely used in various areas such as finance, marketing, communication, web service, surveillance and security. The continuing growth in computing hardware and consumer demand has led to a rapid increase of multimedia data searching. With the rapid development of computer vision and communication techniques, real-time multimedia data mining is becoming increasingly prevalent. A motivating application is Closed-Circuit Television (CCTV) surveillance systems. However, most data mining systems mainly concentrate on text based data because of the relative mature techniques available, which are not suitable for CCTV systems. Currently, CCTV systems rely heavily on human beings to monitor screens physically. An emerging problem is that with thousands of cameras installed, it is uneconomical and impractical to hire the required numbers of people for monitoring. An Intelligent CCTV (ICCTV) system is thus required for automatically or semi-automatically monitoring the cameras.

BACKGROUND

CCTV Surveillance Systems

In recent years, the use of CCTV for surveillance has grown to an unprecedented level. Especially after the 2005 London bombings and the 2001 terrorist attack in New York, video surveillance has become part of everyday life. Hundreds of thousands of cameras have been installed in public areas all over the world, in places such as train stations, airports, car parks, Automatic Teller Machines (ATMs), vending machines and taxis. Based on the number of CCTV units on Putney High Street, it is "guesstimated" (McCahill & Norris 2002) that there are around 500,000 CCTV cameras in the London area alone and 4,000,000 cameras in the UK. This suggests that in the UK there is approximately one camera for every 14 people. However, currently there is no efficient system to fully utilize the capacity of such a huge CCTV network. Most CCTV systems rely on humans to physically monitor screens or review the stored videos. This is inefficient and makes proactive surveillance impractical. The fact that police only found activities of terrorists from the recorded videos after the attacks happened in London and New York shows that existing surveillance systems, which depend on human monitoring, are neither reliable nor timely. The need for fully automatic surveillance is pressing.

Challenges of Automatic Face Recognition on ICCTV Systems

Human tracking and face recognition is one of the key requirements for ICCTV systems. Most of the research on face recognition focuses on high quality still face images and achieves quite good results. However, automatic face recognition under CCTV conditions is still on-going research and many problems still need to be solved before it can approach the capability of the human perception system. Face recognition on CCTV is much more challenging. First, image quality of CCTV cameras is normally low. The resolution of CCTV cameras is not as high as for still cameras and the noise levels are generally higher. Second, the environment control of CCTV cameras is limited, which introduces large variations in illumination and the viewing angle of faces. Third, there is generally a strict timing requirement for CCTV surveillance systems. Such a system should be able to perform in near real-time — detecting faces, normalizing the face images, and recognizing them.

MAIN FOCUS

Face Detection

Face detection is a necessary first step in all of the face processing systems and its performance can severely influence on the overall performance of recognition. Three main approaches are proposed for face detection: feature based, image based, and template matching.

Feature based approaches attempt to utilize some priori knowledge of human face characteristics and detect those representative features such as edges, texture, colour or motion. Edge features have been applied in face detection from the beginning (Colmenarez & Huang 1996), and several variations have been developed (Froba & Kublbeck 2002; Suzuki & Shibata 2004). Edge detection is a necessary first step for edge representation. Two edge operators that are commonly used are the Sobel Operator and Marr-Hildreth operator. Edge features can be easily detected with a very short time but are not robust for face detection in complex environments. Others have proposed texture-based approaches by detecting local facial features such as pupils, lips and eyebrows based on an observation that they are normally darker than the regions around them (Hao & Wang 2002). Color feature based face detection is derived from the fact that the skin color of different humans (even from different races) cluster very closely. Several color models are normally used, including RGB (Satoh, Nakamura & Kanade 1999), normalized RGB (Sun, Huang & Wu1998), HSI (Lee, Kim & Park 1996), YIQ (Wei & Sethi 1999), YES (Saber & Tekalp 1996), and YUV (Marques & Vilaplana 2000). In these color models, HSI is shown to be a very suitable when there is a large variation in feature colors in facial areas such as the eyes, eyebrows, and lips. Motion information is appropriate to detect faces or heads when video sequences are available (Espinosa-Duro, Faundez-Zanuy & Ortega 2004; Deng, Su, Zhou & Fu 2005). Normally frame difference analysis or moving image contour estimation is applied for face region segmentation. Recently, researchers tend to focus more on multiple feature methods which combine shape analysis, color segmentation, and motion information to locate or detect faces (Qian & Li 2000; Widjojo & Yow 2002).

The Template matching approach can be further divided into two classes: feature searching and face models. Feature searching techniques first detect the prominent facial features, such as eyes, nose, mouth, then use knowledge of face geometry to verify the existence of a face by searching for less prominent facial features (Jeng, Liao, Liu & Chern 1996). Deformable templates are generally used for face models for face detection. Yuille et al. (1989) extends the snake technique to describe features such as eyes and the mouth by a parameterized template. The snake energy comprises a combination of valley, edge, image brightness, peak, and internal energy. In Cootes and Taylor's work (1996), a point distributed model is described by a set of labeled points and Principal Component Analysis is used to define a deformable model.

Image-based approaches treat face detection as a two class pattern recognition problem and avoid using *a priori* face knowledge. It uses positive and negative samples to train a face/non-face classifier. Various pattern classification methods are used, including Eigenfaces (Wong, Lam, Siu, & Tse 2001), Neural Network (Tivive & Bouzerdoum 2004), Support Vector Machine (Shih & Liu 2005), and Adaboost (Hayashi & Hasegawa 2006).

In summary, there are many varieties of face detection methods and to choose a suitable method is heavily application dependent. Figure 1 shows various face detection techniques and their categories. Generally speaking, feature-based methods are often used in real-time systems when color, motion, or texture information is available. Template-matching and image-based approach can attain superior detection performance than feature-based method, but most of the algorithms are computationally expensive and are difficult to apply in a real-time system.

Pose Invariant Face Recognition

Pose invariant face recognition can be classified into two categories: 2D based approaches and 3D based approaches. Although 3D face models can be used to describe the appearance of a human face under different pose changes accurately and can attain good recognition results for face images with pose variation, there