Chapter 1

Introduction

If it were possible to detect and track human hands in video sequences, a variety of useful applications would be possible. These applications include human-computer interaction, human-robot interaction, gesture and sign-language recognition, intelligent security systems and more.

Over the last 15 years, the problem of hand tracking has become an attractive area for research in the field of computer vision. Many early hand tracking systems relied on uncluttered static backgrounds, high resolution imagery, and manual initialization. Most of the modern hand tracking systems are oriented towards sign language recognition, human-computer interaction, and human-robot interaction. In these applications, it is possible to make the very useful assumption that only hands are moving while the rest of the scene is stationary. The problem can be further simplified by assuming that there will be only two hands, since there should be only one person performing sign language or gestures in the scene. Nowadays, the systems are becoming more robust, but they generally still require high resolution imagery.

1.1 Problem Statement

We are primarily interested in hand detection and tracking because monitoring peoples' hands could be a key to predict what that person is doing. In security applications, it would be very useful to detect and track hands of people in the scene and perform automated analysis of their actions, e.g., by determining if they are walking, running, punching someone, or even identifying any object they are holding. Detecting and tracking hands in security applications is more challenging than in other applications such as human-computer interaction because most surveillance cameras provide noisy images, with human figures quite far away and therefore appearing at a fairly low resolution. The resolution of hands in those images may be as small as 24×24 pixels or even smaller; detecting such small hands in static images is a very challenging task. Another difficulty is that motion information is possible in the security application which aim is to locate human hands and identify the object in the hand from single image. The problem of detecting hands in single image becomes more difficult if the hands in the image are in low resolution and noisy.

To track hands in video sequences, we must first *detect* hands, then use some tracking algorithm to link detected hands from frame to frame. The majority of the hand trackers rely mostly on relatively simple detection algorithms to initialize the tracker, then fairly powerful tracking algorithms maintain an estimate of the state of that hand. For example, some

systems simply use a skin detector to detect skin blobs. Systems using such simple detection algorithms will be less robust when video sequences are cluttered with many potential hand blobs. A system that is able to robustly detect hands in static images will be a major contribution to the development of robust hand tracking systems. The hand bounded by the square in Figure 1.1 is indeed the size of 24×24 pixels and is corrupted by noise. When the bounded region is cropped out and zoomed as show in lower right corner of Figure 1.1, it is difficult even for humans to recognize what it is. The ultimate goal of my research is to construct a system which is able to detect multiple small hands like this in cluttered noisy single images.

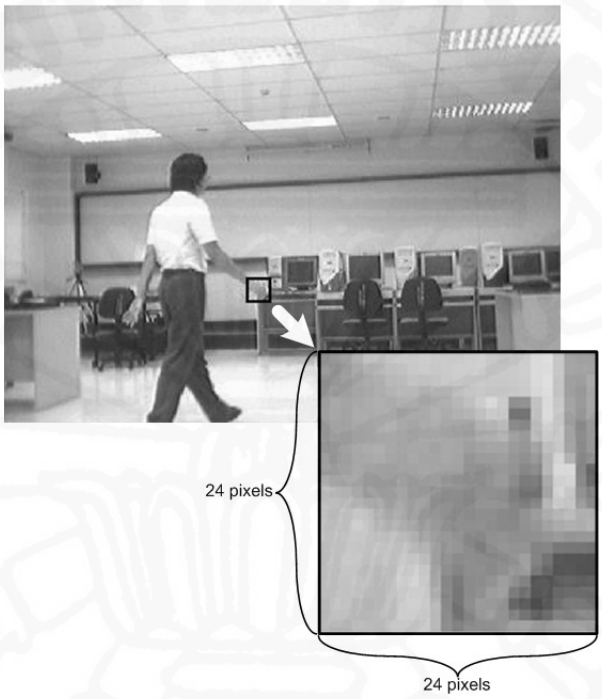

Figure 1.1: Low resolution and relatively noisy image (640×488) captured inexpensive IEEE1394 web cam. By looking at zoomed hand, it is obvious that detecting hands in this image is very challenging.

1.2 Approach In This Thesis

Recently, several face detectors [1, 2, 3] have been introduced and some face detectors are now in commercial applications such as focusing on faces in digital cameras. Among those face detectors, I hypothesized that the Viola and Jones robust real-time face detection cascade [3] would be useful for detecting hands in static images. But detecting low-resolution hands is much more challenging than detecting faces since faces have are well structured, with mouth, eyes, eyebrows, and nose in predictable positions. This structural information is available even on faces in low-resolution noisy images. For low-resolution hands, in contrast to faces, there is little useful structural information except the contour of hand, as shown in Figure 1.1. I found that a straightforward detector based on the Viola and Jones cascade [3] is not sufficient but does help to eliminate more than 95% of false positives at very high speed. To provide better performance, I utilize other useful features of hands like skin color and geometric properties to eliminate the remaining 5% of false positives.

I have conducted a thorough evaluation on our proposed system and found that its detection rate was 86.8% and that its false positive rate was 1.19 false detections per image on average. The system's speed and accuracy will enable many useful applications that are based on hand detection and tracking.

1.3 Organization Of This Thesis

This thesis is organized with six chapters:

Chapter 1 Brief introduction to hand detection and tracking, its applications, and an overview of the thesis.

Chapter 2 Literature review on hand detection and tracking not only in modern days but also in early time.

Chapter 3 Detailed explanation of the system architecture and each building block in my hand detection system.

Chapter 4 Thorough step-by-step description of how data acquisition, training and testing are carried out.

Chapter 5 Presentation and discussion of the results of the experiments described in Chapter 4.

Chapter 6 Conclusion and recommendations for those who would like to use or improve upon this work.