CHAPTER 1

INTRODUCTION

1.1. The subject of this thesis

The main subject of this thesis is *motion* – especially the study and automated analysis of the different motion patterns (*gestures*) which are revealed in a sequence of movements performed by the human body as a means of communication either with other human beings, or with a computer. Motion is arguably the most important source of information in biological vision systems. Many lower animals are blind to anything that is not moving, and in the more sophisticated nervous systems of the higher vertebrates, including man, motion stimuli are usually the first to which attention is paid. It is well-established fact that the visual process does not provide much information about retinal images unless they are moving or changing in some way. In many cases, the percept corresponding to a stabilized, stationary retinal image will fade and eventually disappear. This is in striking contrast to the instantaneous "pop-out" of any moving or changing stimuli. For living creatures, the importance of possessing a system for effi-

cient processing of motion information is not surprising, having in mind that biological organisms have evolved and inhabit an environment which is alive and ever-changing, and where acting in accordance with the unceasing changes is often essential for survival. In the artificial world of computers, until recently visual capabilities were not regarded as essential, but this attitude is now rapidly changing as technological advancement has finally made it possible to implant more intelligence into a hardware which is getting faster, cheaper and more powerful from day to day. It has become clear that if the next generation of "personal robots" are to be more human-like and to change their working place from the factory floor to the office or private home, they will need a visual system and will have to process motion stimuli to "survive" and fulfill their functions in these new environments. This will require the development of new basic technologies utilizing visual pattern recognition methods and our main objective here will be to propose a gesture recognition system which could be used as one of the building blocks of a more sophisticated visual interface between humans and computers.

1.2. Motion recognition applications

Because of the richness of the information contained in changing motion patterns, motion recognition and analysis have found application in many different areas:

- in TV communications, video conferencing to exploit temporal redundancy to reduce transmission rate;
- in mobile robotics to navigate autonomously in changing or unknown environments;
- in satellite imagery and meteorology to measure cloud motion and establish wind maps;
- in military applications for target tracking and autonomous navigation;

- in biomedical imagery for heart motion, human motion (in sports, reeducation) analysis;
- in surveillance and monitoring for intrusion detection, road traffic analysis;
- in human-computer interfaces and virtual reality for analyzing facial motion, lip motion, human gestures;
- in computer vision functions for segmentation of images, tracking, prediction of environmental layout, recovery of depth and relative motion viewing system and environment.

Among those mentioned above, of primary interest to us here will be the area which takes interest in human-computer interfaces (HCI). The recent increase in computational power and storage capacity of personal computers, together with the availability of image acquisition devices at reasonable prices, have led to an increased interest in the creation of systems capable to provide more refined HCIs (see Pavlovic et al., 1997, for a review on the use of hand gestures for HCI). Considering the importance of visual information for humans, gesture recognition will necessarily be a major component of such interfaces. For a successful user-independent gesture recognition system, good generalization abilities are essential, and for this end it has to be provided with the following features:

(a) to be robust to changes in background and illumination conditions;
(b) independence to subjects' external appearance (including gender, body size, clothing, etc.);
(c) ability to cope with the non-uniformity in the speed of the gestures;
(d) to be robust to shifts in subjects' spatial position, both in the horizontal plane and in depth.
(e) if the system is to be used as part of a human-computer interface, *real time* performance is indispensable;

(f) it is desirable that all processing stages are automated (i.e. no manual processing is involved) and the users of the system are not burdened with cumbersome special equipment like hand gloves, markers, sensors, etc.

Although methods which deal only with static gestures or posture recognition also exist, for the recognition of dynamic actions such as gestures, it seems more natural and efficient to attempt some form of motion extraction/estimation from the raw input data, before proceeding to motion recognition. Many different methods for motion evaluation (Horn and Schunck, 1981; Hildreth, 1984; Adelson and Bergen, 1985; Heeger, 1987; see Mitiche and Bouthemy, 1996 for a survey; Barron et al., 1994, for performance comparisons) and motion-based recognition (e.g. Davis and Shah, 1994; see Cedras and Shah, 1995 for a survey) have been proposed. However, most of those methods deal predominantly with what is known in the psycho-physical literature as *common motion* (this and some other related concepts will be explained in more detail in chapter 3). Although it has been shown by Cutting and Proffitt (1982) that in the case of human motion, the *relative motion* between the elements of an object might be more informative for its recognition than *common motion*, the possible contribution of relative motion to motion recognition has not yet been explored adequately.

1.3. Outline of the approach

In this thesis we propose a method for user-independent real-time gesture recognition from time-varying image sequences based on the following approach: we do not assume any fixed model for the moving body parts, but we rather extract features about local motion changes across the whole dynamic scene, and combine them in short predicative primitives, containing information about the patterns of motion changes in many directions *relative* to certain reference points. This bottom-up approach for extraction of "relative motion" dependent primitives in the early processing stage of our system, is combined with top-down based learning and noise-filtering abilities at the later stages.

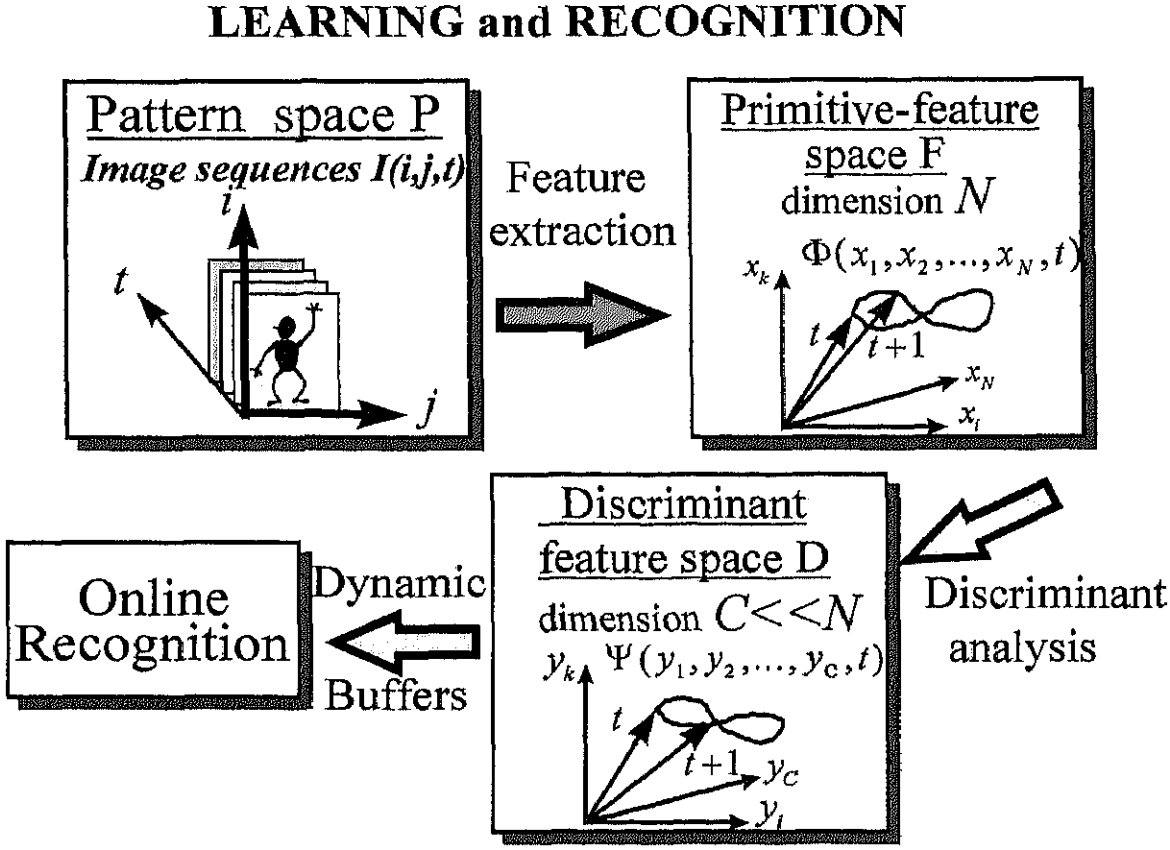

Figure 1-1. General outline of the system.

The method proposed here operates in three stages, shown in Fig.1-1, and each of these stages will be the subject of a separate chapter (chapters 3-5). In the first stage (*primitive features extraction*), a set of primitive features related to relative motion dependent binary pattern changes are extracted from the input image sequence. Thereby, input data are mapped from pattern space ***P*** to primitive feature space ***F***. In the second stage (*learning*), the primitive features extracted at the first stage are linearly combined on the basis of multivariate analysis, using linear discriminant analysis (LDA) (Fisher, 1936; Duda and Hart, 1973; Otsu and Kurita, 1988) to provide new and more efficient

features. The learning process determines the mapping from primitive feature space $\boldsymbol{F}$ into discriminant feature space $\boldsymbol{D}$, where the different classes of gestures depict different trajectories. Recognition is performed by comparing a test-sample gesture's trajectory in $\boldsymbol{D}$ to all class-representative trajectories of previously learned gestures, and classifying it to that class which is most similar in terms of a certain distance measure. The function of the *dynamic buffers structure* (DBS), which forms the final stage of our system, is to provide an online segmentation of the test gesture sequences and at the same time to filter out some of the noise present in the output from the LDA-based classifier. To evaluate the performance of the proposed method it has been tested with several different data sets, some of which have been created to incorporate the requirements for generalization abilities mentioned above.