CHAPTER 2

Related Work

Over the last two decades there has been significant interest in the computer vision community in *motion extraction* and *motion-based recognition*, which are the primary concern for us in this thesis. Many different methods have been proposed and new methods continue to appear, so a detailed survey of the area would be beyond the scope of the present thesis (see Barron et al., 1994; Cedras and Shah, 1995; Mitiche and Bouthemy, 1996; etc. for good reviews). However, some of the relevant concepts and research effort will be briefly presented in this chapter.

2.1. Methods for motion extraction

Although various methods exist for two-dimensional motion extraction, they can be grouped into three main groups: optical flow, motion correspondence (tracking) and region-based features. These will be briefly explained below.

Optical flow is an approximation of the two-dimensional flow field from image intensities. Many methods have been developed (see Barron et al., 1994 for a review), however, accurate and dense measurements are difficult to achieve. Optical flow methods can be divided into four groups:

(a) differential methods – velocity is computed from spatio-temporal derivatives of image intensities. Methods for the computation of first-order (Horn and Schunck, 1981; Lucas and Kanade, 1984) and second order derivatives (Nagel, 1983) have been devised, although estimates from second order approaches are usually poor and sparse;

(b) region based matching – the velocity is defined as the shift yielding the best fit between image regions, according to some similarity or distance measure (e.g. Anandan, 1989);

(c) energy-based methods – optical flow is computed using the output from the energy of velocity tuned filters in the Fourier domain (e.g. Adelson and Bergen, 1986; Heeger, 1988);

(d) phase-based methods – velocity is defined in terms of the phase behavior of band-pass filter outputs, for example the zero-crossing techniques (e.g. Fleet and Jepson, 1990).

Problems with optical flow methods include being susceptible to the aperture problem, which in some conditions only allows the precise computation of the normal flow, i.e. the component parallel to the gradient. They are also prone to boundary oversmoothing, have problems with multiple moving objects, where segmentation can be difficult to achieve, etc.

Motion correspondence is concerned with the matching of characteristic features (tokens) through time from which motion trajectories can be generated and analyzed subsequently (e.g. Rangarajan and Shah, 1991). The tokens need to be distinctive enough for easy detection and stable enough through time so that they can be tracked. Tokens include edges, corners, interest points, regions, body parts, etc. This method is

generally good in the case of rigid motion, however problems arise when non-rigid motion objects have to tracked through multiple frames.

Region-based features are features generated from the use of information over a relatively large region or over whole images. For certain types of objects or motions, the extraction of precise motion information for each single point is neither desirable nor necessary. Instead, the ability to have a more general idea about the content of a frame might be sufficient. The approach which will be presented in the next chapter belongs to this method type (another region based methods are e.g. Polana and Nelson,1992, the mesh features in Yamato et al.,1992; the "eigenlips" in Kirby et al., 1993, etc.). The problem with region-based features is that they are usually more abstract and their interpretation might not be always obvious.

2.2. The study of gestures

Defining gestures is not an easy task. In the Webster's dictionary gestures are defined as "... the use of motions of the limbs or body as a means of expression; a movement usually of the body or limbs that expresses or emphasizes an idea, sentiment or attitude." In the psychological and social studies the definition is more related to man's expression and social interaction. In the area of HCI, which will be of primary interest for us here, more practical meaning is attached to gestures, which are usually interpreted as a means to control or communicate with computers by the use of the human body. The final goal is to interpret the contents of real-time image sequences using vision based recognition techniques. However, before starting to create and apply computer vision procedures to analyze gestures, it might be helpful to have a closer look at gestures, with the idea that knowing what types of gestures exist and what are their peculiarities might be useful when trying to design primitives and algorithms for their processing.

Several alternative taxonomies of gestures have been developed in the literature. For example, Kendon (1986) distinguishes "autonomous gestures" which occur independently of speech, from "gesticulations" which occur in association with speech. McNeill and Levy (1982) recognize three groups of gestures: iconic, metaphoric and "beats". From the viewpoint of HCIs a different taxonomy has been developed by Quek (1994, 1995) and it will be briefly reviewed here. First, all movements are classified into two major classes: *gestures* and *unintentional movements*. *Unintentional movements* are movements which do not convey any meaningful information, while *gestures* can be divided into two types: *communicative* and *manipulative*. *Manipulative gestures* are those used to act on objects in the environment and usually are not necessarily visually interpretable, which renders them less important from the viewpoint of visual HCIs. *Communicative gestures*, on the other hand, have direct communicational purpose and are usually intended for visual interpretation. The difference between these two groups can be illustrated by the following two examples: an orchestral conductor's hand motions are communicative gestures intended to communicate interpretative information to the orchestra, while a pianist's hand movements can be considered manipulative gestures which would be difficult (and essentially are not intended) for visual interpretation. A glove device or some similar sensor would be much more suitable if hand motion has to be interpreted in the latter case. Communicative gestures are further divided into *symbols* and *acts*. *Symbols* are gestures that have a linguistic role. They can be *referential* or *modalizing*. The former operate independently to designate objects or concepts, e.g. rubbing one's index finger and the thumb is referential to money (the referent). *Modalizing* gestures usually operate in conjunction with other means of communication, e.g. speech, to indicate the opinion of the speaker. Extending one's hand apart while speaking about how big a fish has one caught is an example of modalizing gesture and in this case the meaning of the conversation would be unclear if only the audio information was analyzed. *Acts* are gestures that are directly related to the interpretation of the movement itself and can be *mimetic* (imitating some action) or *deictic* (pointing

acts). Deictic gestures are very suitable for use in computer input which renders them of primary importance for HCIs. Deictic gestures can be further classified into *specific* (when the subject selects a particular object or location), *generic* (eliciting the identity of a class of an object by picking one of its members) and *metonymic* (pointing to an object to signify some entity related to it, e.g. pointing to a picture of skyscrapers to signify New York city).

The temporal characteristics of gestures represent another important issue which might be useful to segment gestures from other unintentional movements. It has been established that a gesture (or a "gesture phrase") generally consists of three phases: *preparation*, *nucleus* (also called peak or stroke), and *retraction*. The *preparation* phase consists of a preparatory movement that sets the hand in motion from some resting position, while the *retraction* phase returns the hand to rest or reorients it for a new gesture. The *nucleus* of a gesture has some definite form and enhanced dynamic qualities, and is easily differentiable from the other two phases.

2.3. Motion-based gesture recognition

Motion-based recognition is an approach that favors the direct use of motion information extracted from a sequence of images for the purpose of recognition. Although many different methods exist, usually two main steps are used. The first step consists of finding an appropriate representation for the objects or motions to be modeled from the motion in the image sequence. These representations can be either lower level bottom-up features, or organized into very high-level representations like the motion verbs described by Koller et al. (1991) and Tsotsos (1980). The second step consists of the matching of some unknown input with a model. The methods here are more standard, and are often common pattern classification techniques. Several representative motion-based recognition methods for the case of human movements recognition, which is of main interest for us here, will be briefly reviewed below.

One of the most successful methods for human motion recognition has been proposed by Yamato et al. (1992), who take a probabilistic approach to the classification of motion using Hidden Markov Models (HMM). A HMM consists of a set of internal "hidden" states Q, a set of output symbols V, a matrix A whose elements consist of probabilities of transition between every state, a matrix B of output symbols probabilities for each state, and a vector π of initial state probabilities. The HMM changes from state j to state k with probability $A(j, k)$. An image sequence is processed in three steps. First, an observed sequence O of output symbols is derived using mesh features. In the second step, sequences are used to train the HMMs, and there are as many HMMs as there are different motions. The parameters of each model are optimized for a certain motion pattern using the Baum-Welch algorithm (Yamato et al., 1992). Finally, the recognition is done by observing a certain output sequence O and finding the HMM which is most likely to generate the same sequence.

Another popular method for learning, tracking and recognizing human gestures has been proposed by Darrell and Pentland (1993). The method uses an automatic view-based approach to build the set of view models from which gesture models will be created. The model views of an object are built using normalized correlation. The first view is chosen by the user as one of the images from a sequence. The object in the subsequent input images is tracked and when the correlation score drops below a certain threshold a new model view is created with the current input image. This process is repeated until no more models are necessary. Once all views of an object have been gathered gesture models are created as a set of views over time. The gesture models are first adjusted to the same length using dynamic time warping (DTW) (Sakoe and Chiba, 1979). To compare a new input gesture, each frame of the new sequence is correlated with a model view and a score is determined. The score result for the whole sequence is plotted with respect to time. The same process repeated for all model views and the score results for each model are stored in a vector. The input gesture is compared to all gesture models and the peaks in the plotted scores indicate a match.

Another line of research, known as 2D or 3D *model-based* methods (in contrast to the *appearance-based* methods described above) approaches the problem of gesture recognition by first creating a model of the posture and motion of the hand or the human body, after which gestures are inferred from the estimated model parameters (e.g. joint angles, palm positions, etc.). The models in this group can be divided into *volumetric* models and *skeletal* models. *Volumetric* models are meant to describe the 3D visual appearance of the human hand, arms or body (O'Rourke and Badler, 1980; Magnenat-Thalmann and Thalmann, 1990; Downton, 1991;Etoh et al., 1991; Koch, 1993; Wren et al., 1996). They are usually used for analysis-by-synthesis tracking and recognition of the body's posture, i.e. body's posture is analyzed by first synthesizing a 3D model of the body after which its parameters are varied until the model and the real human body appear as the same visual images, or structures like generalized cylinders and super-quadrics are used to approximate the shape of the different body parts like finger links, forearm, upperarm, etc. Possible problems with this approach are that the dimensionality of parameter space is high for the more elaborate models, and more importantly, obtaining the necessary parameters in a reliable and robust manner via computer vision techniques might be difficult. In the *skeletal* models (Lee and Kunii, 1995; Vaillant and Darmon, 1995), instead of dealing with all parameters as in the volumetric models, models with a reduced set of equivalent joint angle parameters together with segment lengths are used and morphology and biomechanics based constraints are imposed on the parameters.