A Survey of Vision-based HCI

Beifang Yi[1]

February 11, 2003

[1]Email: b_yi@cs.unr.edu

Abstract

Abstract

This report is a brief survey of current development of computer vision based researches on human-computer interfaces. The emphasis is put on the the studies of eye/pupil gaze. The other areas include face pose, head gestures and hand gestures. The analysis of multimodal HCI is also discussed. At the end of the paper we list some of the conferences, journals, websites, and academic groups that are related to the CV-based HCI.

Contents

0.1 Introduction

Human-Computer Interaction/Interface (**HCI**), a relatively new but vigorously developed discipline, has not only evolved as an active research area with the rapid development of computer science and engineering technologies but also gained new substances from the results of other subjects such as psychology, sociology, anthropology, and industrial design. During passing years a significant number of academic institutions and industrial corporations have paid more and more attention on HCI — the ease-of-use environment between man and machine. Many various research topics have introduced the concepts of HCI and more and more industrial products that have no relation with one another are making use of the methods of HCI. Different researchers and users have different viewers on HCI or even different definitions of HCI. In this section we first introduce one or two definitions representative of HCI and the research areas concerned with HCI, and then focus on the current development of computer-vision-based HCI.

0.1.1 About HCI

To give a definition for a term of a scientific discipline, it always is a good idea first to look at what research areas the discipline has emphasized on. As for HCI, Andrew Sears [132] introduced three definitions by other scholars for HCI in his understanding of HCI: 1)."Human-Computer Interaction (HCI) is about designing computer systems that support people so that they can carry out their activities productively and safely." (by Preece et al); 2). "Human-Computer Interaction (or HCI) is, put simply, the study of people, computer technology and the ways these influence each other. We study HCI to determine how we can make this computer technology more usable by people." (by Dix et al); and 3). "Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them." (by Hewett et al). From the above definitions, four important ideas stand out: people, computing systems, interaction, and usability.

The term *people* here is referred to as an individual or a group of people working together to solve a problem [132]. Although many people take the implication of *User*-centered interface when mentioning of HCI, Miller [114] prefers to use *Human*-computer interface instead of *user* interface for three

reasons, the most important one of which is that the domain of user and the user population have been changed rapidly. Here *people* and *human* are the same in the definition. *Computing systems* differ from traditional *computers*. Computer vision, computer graphics, image processing, operating system, and informatics have been combined to the system of computing. *Usability* concerns with a system's easiness to learn, efficiency and satisfactoriness to use. Researchers evaluate the usability of a system they have developed by implementing and testing the design. Finally, *interaction or interface* is the key word in HCI. Hewett et al [92] describe HCI as an interdisciplinary area with emphasis on the "joint performance of tasks by humans and machines", "the structure of communication between human and machines". For example, in the domain of computer science, the *interaction* takes the form of "application design and engineering of human interfaces", in psychology, that of "application of theories of cognitive processes and the empirical analysis of user behavior", in sociology and anthropology, that of "interactions between technology, work, and organization",and in industrial design, that of "interactive products.

Thus HCI "studies a human and a machine in communication" and "draws from supporting knowledge on both the machine and the human side [92]." On the human side, cognitive psychology has always concerned with the studies of human information processing and performance and of the learning of HCI system. For example, Pelz and his group focus on the visual perception in everyday life by examining the eye movements of subjects as they perform complex tasks in natural environment, to explore the process on how human stores and recovers the information and how human uses that information in planning and guiding actions [17]. Hewett et al proposed [92] that "it is important to understand something about human information-processing characteristics, how human action is structured" such as models of cognitive architecture and human action, phenomena and theories of attention, vigilance, motivation, and learning and skill acquisition, about "the nature of human communication" such as aspects of natural languages, specialized languages", and about "anthropometric and physiological characteristics of people and their relationship to workspace and environmental parameters" such as human cognitive and sensory limits, display technologies and design, fatigue and health issues, and design for stressful environments and for the disabled.

On the other side, some specialized machine (computer) components play

an important role in interacting with humans. Several topics connected with these components are covered in [92]: 1). input and output devices including virtual devices; 2). dialogue techniques (dialogue input/output, interaction techniques); 3). dialogue genre(style and esthetics, interaction metaphors); 4). computer graphics(2/3-D geometry, graphics primitives and attributes, solid modeling, and color representation); and 5). dialogue architecture(multi-user/look-and-feel, screen imaging models, window manager models). Myers considered as the interactions in HCI those important innovations in computer science such as direct manipulation(visible objects on the screen directly manipulated with a pointing device), mouse pointing device, multiple tiled windows, drawing programs, text editing, spreadsheets, HyperText, CAD, Multi-Media, 3D system, VR(Virtual Reality), AR(Augmented Reality), and natural language and speech [117].

0.1.2 Modalities of HCI

Keyboard and mouse have been the basic and the most common devices in the information exchanges between human and machine in the traditional computer systems. With the coming of HCI and its growing more and more fledged, new modalities for HCI have received received great attention and are being developed. HCI is now trying to take as its modalities the the natural means that human employ to communicate with each other: vision, hearing, touch, smell, and taste. As Sharma et al pointed out [134], "Almost any natural communication among humans involves *multiple, concurrent modes of communication*", therefore "people prefer to interact multimodally with computers". They considered four basic questions about multiple-modalities for HCI: 1). *why* integrate multiple modalities; 2). *which* modalities to integrate; 3). *when* to integrate multiple modalities; and 4). *how* to integrate multiple modalities. Here we only introduce the first two topics. As an answer to the first question, they proposed three reasons: a). practical reasons: the mouse-keyboard-based system are unnatural and cumbersome; the current advanced single-modality HCI lacks robustness nab accuracy; and multiple modalities (particularly with redundant input) in HCI allow physically or cognitively handicapped people access to to computers. b). biological reasons: the integration of multiple sensory modalities is the most common thing in the natural world. c). mathematical reasons: the optimal ways of integrating

different sensory data can produce best detection rates in the area of target detection. And a system using single modality may not reduce the uncertainty for decision making. Furthermore, "it is statistically advantageous to combine multiple observations from the same source because improved estimates are obtained using redundant observations" and "multiple types of sensors may increase the accuracy with which a quantity can be observed."

They discussed the second question under the two categories: *human-action modalities* and *computer-sensing modalities*. In the former case, *hand movements* has been the most exploited modality because of the dexterity of the human hand. The corresponding interface devices are keyboard, mouse, stylus, pen, magnetic wand, joystick and trackball. The use of *hand gestures* is the next one of human-action modalities: from simple pointing through manipulative gestures to more complex symbolic gestures (for example, American Sign Language). Here we have the relevant interface glove-based or video-camera devices. Another dominant human-action modality is the production of sound, especially spoken words, which is deeply connected with one visible action—lip movement. Recently, eye movements, facial expression, and body motion have received more and more attention.

In the case of computer-sensing modalities, five kinds of modalities have been proposed. 1). Position and motion sensing: keyboard for typing input; mouse, light pen, stylus, and tablet for 2D plane input; joystick and trackball for 3D sensing; and many other advanced devices (particularly, those in the field of computer vision) for locating and tracking the eye/gaze, head/pose, hand/fingers, and even the body movements. 2). Audio sensing: microphones are used to sense the sound waves and the techniques of automatic speech recognition can interpret speech, the most natural human-action modality for HCI. It is well known that the visual sensing modality of lip motion can improve the recognition rate for speech [149] [150]. 3). Tactile and force sensing: force sensing by using appropriate haptic devices plays an important role in "building a proper feel of *realism* in virtual reality". 4). Neural sensing: monitoring of brain EEG activities can give one more computer-sensing modality. Also this brain electrical impulses can be used for brain-activated control (BAC). It is particularly promising in the field of aircraft piloting and for the disabled. 5). Visual sensing: so many and various human-action modalities can be incorporated into HCI just by using a video together with a set of techniques. These modalities include the very areas of computer vision: hand gestures, lip movement, eye/gaze

tracking, facial expressions, head pose, and other body movements. In the following sections we will discuss in detail the current development of those computer vision topics and how those modalities are integrated into HCI context.

0.1.3 Vision-based HCI

By vision we refer to the use of cameras and a set of visual or graphical techniques for presenting and processing information. Compared to the non-vision-based HCI methods, vision-based HCI has the advantage of unobtrusiveness and gives a sense of "naturalness" and being comfortable during the process of the human-machine interactions. For example, glove-based gestural$\Phi\Phi\Phi$ interfaces for hand gesture recognition require that the user put on a cumbersome device and carry cables connecting the device to a computer, thus depriving the user of the ease and naturalness with which he/she interacts with the computer-controlled environment.

But for the present time, use of visual sensing for HCI is incorporated as a vigorous and promising method into the other modalities of HCI. For example, in a system for detecting physiological features of emotional stress during driving, a skin conductance sensor, a respiration sensor, blood volume pulse sensor, and an electrocardiograph are placed on the different parts of the body for detecting skin conductance, respiration, muscle activity and heart activity, while a digital camera is mounted on the steering column to record the facial expressions and actions of the driver and a camera is placed on the dashboard to capture road and traffic conditions. The video signals from the cameras are synchronized with that from the physiological sensors, allowing an unambiguous record of physiological response to driving events [89].

In another example, we can see that computer-vision-based methods and modalities play the most dominant role in a HCI environment. Pinhanez and Bobick have described a computer theater play "*It/I*" [124] which lasts 30 minutes in 4 scenes. This pantomime play has two characters: *It*, a non-human body composed of CG-objects projected on screens, and *I*, a human actor. Three cameras rigged in front of the stage are the input of the computer system, and large back-projected screen, speakers, and stage lights the output. *It* interacts with *I* through images and videos projected on the screens, through sound on stage speakers, and through the stage lights

within *It*'s limit of understanding the world: "the character's reaction is mostly based on tracking *I*'s movements and position and on the recognition of some specific gestures. In other word, "the actions of *I* were restricted to those that the computer recognize through image processing automatically", for instance, sit-down, stand-up, making a certain pose, taking pictures, approaching to an object, and even trying to hang himself. And the *I*'s position can be detected. The computer theater helps to facilitate the construction of interactive, immersive systems where human actor knows how to communicate with computer character and the audience may lean the basic structure and interaction modes the the play.

There are so many areas of research and application where computer-vision-based HCI modalities, with their invasiveness and naturalness, have been combined with other HCI modalities. For example, IVE (Interactive Virtual Environment) interface [157], speech and vision integration system [150] [149], a perceptually-based, interactive, narrative plays pace for children [66], a platform for simultaneously tracking of multiple people and recognition of their behaviors for high-level interpretation [139], 3-D Visual Operating System [1], gaze-assisted translator [96], eye interpretation engine [76], an intelligent mediator [135], affective computing system [126] [123] [122], affection detection, emotion mouse [31], driver's fatigue monitoring system [52] [98].

In the following we concentrate on the vision-based HCI modalities: eye/gaze, head/pose, head gesture, hand gesture, and facial expression.

0.2 Review of Current Research in Gaze

0.2.1 Summary

Cognitive studies and psychological tests have demonstrated that we humans sometimes are not consciously aware of our own activities, particularly, the eye movements in our everyday lives [131] [17] [143]. For example, a study of natural eye actions in a task under natural environment tells us that there exist *planful* eye movements (without a subject's consciousness) to objects well in advance of the subject's interaction with the object and very short fixation durations in all tasks, some as short as 33 msec [17]. Another group [76], in the process of developing an eye-aware application tool, examined and cat-

egorized the *natural* eye-movement behavior and intentions, concluding that two patterns of eye movement: *revisits* and *significant fixation* complement eye's other movements: fixation, saccade, and blink, and thus making use of a new concept **SFT: Significant Fixation Threshold** in their development of algorithms. There are other studies on analyzing and synthesizing the fixation/dwell time, saccades[128], [39]. A research on eye/hand coordination patterns suggests that faster input devices (more than direct hand pointing) can be designed to take advantage of the non-linear transformation in input devices and various hand-eye coordination patterns found for computer target acquisition [143].

Having understood the mechanisms of various patterns of eye movement (**EM**) will deliver guidance to our studies on the CV-based HCI areas of gaze tracking and application. For example, how to get rid of the "unconscious" fixations which the computer **does** detect but are of no importance to the reader; the knowledge of the variation of dwell time (first-time computer users and experienced computer readers differ greatly in their reading speed) is a key point to the dynamic software design.

Based on the latest studies on gaze fixation and tracking, more advanced *eye-aware* application tools/environments are under development in order to understand user behaviors by modeling typical EM, by analyzing and synthesizing different patterns of EM, not merely by using the direct first-step EM (in other words, by extracting and utilizing the *semantics* of EM, not limited to the *grammars* of EM), and thus to control computer with eyes [39], [96], [76]. Gaze studies in computer vision HCI play a very important role. For example, in a *3-D Visual Operating System* [39] where the gaze is used to interpret the user's intention for non-command interactions in the process of human-machine interactions; in a *gaze-assisted translator* [96],which will work this way: when a user reads an on-screen document in a certain foreign language and encounters difficulties (e.g. new words/phrases), the translator will detect such circumstance and deliver necessary assistance (e.g. the corresponding words/phrases in native language) in real-time; in an *eye interpretation engine* [76] that will distinguish, recognize, and adapt to various EM behaviors for different individuals, or different EMs for same person such as looking at screen with/without intentions.

We will first introduce the methodologies for gaze tracking, along with the algorithm, strength for each of them, and then discuss its applications.

0.2.2 Methodologies

All the gazing tracking techniques can be divided as either *IR* (InfraRed) or *Non-IR* approach, and either 2D or 3D tracking, together with various algorithms in its implementation, for example, physiological, geometric method, HMM (Hidden Markov Model), Kalman filtering, neural network, and etc.

InfradRed Technique in Gaze Tracking

InfradRed Technique in Gaze Tracking is to make use of the retro-reflectivity property of the eye reacted to the near infra-red (IR) light in the process of detection and tracking the eyes. Almost all commercial products of pose/gaze tracking introduce this technique and so are many platforms under development. But there are exceptions [141] [48] [51] [39]. The processing speed with IR varies from 30 Hz to 1000 Hz with accuracy of 1.0° to 0.01° [46] [49] [56] [57] [19], whereas that with Non-IR technique, the corresponding specifications are from 15 Hz to 25 Hz with upto 0.4° [39] [48].

Here we only discuss the approaches to the implementation of IR techniques in gaze tracking, because those (which can certainly be combined to the IR techniques) to the realization of Non-IR techniques will be introduced in the subsection *Algorithms*. The retro-reflectivity property of the eye(i.e. the red-eye effect in flash photograph) is exploited for gaze tracking in the IR technique. That is, the near infra red light (for example, a light source of wavelength of 875nm) is almost invisible to human eye but a camera is sensitive to that wavelength. Morimoto et al [115] [80] introduced such a technique that uses two IR time-multiplexed light sources which consist of two rings of eight LED's each: one light source is close to the camera's optical axis (on-axis), the other is placed off-axis. The diameters for both rings are determined empirically in that the inner ring diameter is approximately the same as the camera lens and the the outer diameter is sufficiently large to produce a dark pupil image. The center of the two rings coincide with the camera optical axis and both rings are mounted on the same plane. The even and odd frames of the camera are synchronized with the inner and outer rings's switchings-on respectively so that when the inner ring of illuminators is turned on, an even frame is grabbed and a bright pupil image is produced, and when outer ring is turned on, an odd frame is grabbed and a dark pupil image is caught. The glint (corneal reflection), which is an important factor

for locating the gaze, can be detected from the dark pupil images without the need for stabilizing the pupil brightness because of the concentricity of the both illumination rings with the camera axis. More discussion about this system can also be found at [118].

The eye gaze tracking is accomplished by calculating the subject's pupil and glint locations (glint-pupil vectors) and mapping those locations to the screen coordinates through a brief calibration procedure. The glint on the cornea of the eye can be taken as a reference point and thus the vector from the glint the center of the pupil will give the gaze direction. In the calibration, nine points are arranged in a 3x3 grid on the screen and the user looks at those points in certain order with each point during a fixed period of time. On each fixation of his gaze on one point, the vector from the center of the glint to the center of the pupil at that point is saved and thus gives two second order polynomial equations: one corresponding to x-axis, the other to y-axis. So we can get 18 equations for 12 coefficients of a pair of two second order polynomial equations through which we can calculate the screen coordinates for glint-pupil vectors in the tracking process. The glint-pupil vector can be computed through the difference of the centers of pupil and glint. A window slightly larger than the enclosing box of the pupil is created, the gray scale pixels within are summed horizontally and vertically, and the center of mass of the horizontal and vertical projections determines the x-y coordinates of pupil. Also a search procedure for very bright pixels around the pupil is used to detect the glint and a same method can be used to calculate the center of glint.

The IR technique for tracking eyes, only making use of the geometrical and physiological properties of the eye, does not require geometric models or templates. The scene background becomes irrelevant and the pupils can be detected and tracked in a wide range of scales and illumination conditions compared to other methods for gaze detection and tracking. It also works faster and more accurately.

3D Eye Gaze Tracking

3D gaze tracking, more often, combined with 3D face pose detection and tracking by using either stereo camera pair or 3D algorithm, is a more challenging task [24] [147] [112] [57] [53] [39] [110] [118] [145]. The speed in 3D gaze tracking can reach up to 200 frames per second [57].

Matsumoto and Zelinsky proposed an algorithm for real-time stereo implementation of gaze measurement [113]. The implementation system utilizes two cameras with a field multiplexing device, 3D facial feature models, and 3D eye model which assumes the eyeball to be a sphere. The output video signals from the cameras contains a a stereo image and are multiplexed into one video signal by using the field multiplexing technique. The multiplexed video stream is then fed into a vision processing board, where the direction of the gaze is computed. The 3D eye model consists of three parameters: 1). the relative position (defined as a 3D vector—offset vector from the midpoint of the corners of an eye to the center of the eyeball) of the center of the eyeball respect to the head pose; 2). radius of the eyeball (about 13mm); and 3). radius of the iris (about 7mm). Gaze direction is calculated based on both the pose of the head and the position of the irises of the eye. First, the 3D position of the eyeball can be determined from the head pose. Next, the center of the iris is detected with the circular Hough Transform. Finally, the orientation of the gaze vector is defined by the relationship between the iris center and eyeball center. Since there are four eyes in a stereo image pair, four gaze directions are detected. A single gaze vector is generated by computing the average of those four gaze vectors due to the resolution of the image and the effect of noises. One cycle of the whole tracking process takes about 30ms and the accuracy can reach up to $\pm 1mm$ in translation and $\pm 1^{\circ}$. This system does not require expensive hardwire or artificial environments such as head-mounted cameras, infrared lighting, markings on the face and the 3D coordinates of the features on a face can be directly measured, so it is relatively simple to measure subject's natural behavior.

Liu introduced another contact-free image processing system for the determination of the point of fixation in 3D space [110](point of fixation is defined as the intersection of the gaze line of one eye with the surface of the object glanced at, or the intersection of the two gaze lines for human vision). Two cameras are mounted at the front of the display: the eye camera with a short focal distance for measurement of the 3D eye position, the gaze pan-tilt camera with a long focal distance and with IR LED beside the lens for measurement of gaze function. The eye tracker analyzes the signal from eye camera and detects and tracks the 3D eye position. It finishes two tasks: 1). the face is detected and segmented in 2D video image by using skin color method; and 2). the middle points of both pupils are detected and tracked and the 3D position of one eye is calculated: natural blinks are used to detect

the eye regions, which are registered and stored as reference patterns. The 2D coordinates of both eyes and geometric parameters of the eye camera are exploited to determine the 3D eye position. This 3D information is used to control the pan-tilt gaze camera that the optical axis of the gaze camera follows head movement. The gaze tracker analyzes the signal from the gaze camera, measures the eye movements and evaluates the user's gaze direction. The middle point of pupil and an additional reflection point on the cornea (produced by the Infra LED) determine a feature vector, which is a function of the gaze vector. This function is then determined by a calibration procedure. Then the compensation algorithm, which enables the accurate measurement of the point of fixation without requiring the head being fixed, is exploited to correct the gaze function. Four unified coordinate systems (of display, of eye camera, of gaze camera, and the head-fixed coordinate system) have been created for a set of the head-fixed transformation so that the coordinates of the point of fixation on the display can be calculated. The detailed process can be seen at [110]. This algorithm for the determination of the point of fixation can avoid expensive coordinate transformations in the case of the head movement.The measurement speed is 20 frames per second for the point of fixation and the accuracy is about 0.4°.

Heinzmann and Zelinsky proposed another system for estimating the 3D gaze point, which is done by using a 3D model together with multiple triplet triangulation of feature positions assuming an affine projection [91]. The approach is realized with a 3 layered subsystem. 1). The vision hardware subsystem that tracks a template of an object and results in measured positions. 2). 2D model that receives the measured positions and takes into account geometric constraints in the image plane and the correlation distortion to produce feature positions. 3). 3D model that determines the 3D pose of the head from which 3D point of gaze can be calculated. This system only requires a monocular camera but is able to cope with facial motion in any direction. No close up images are needed. In the event of head movement an active camera is required for compensation of the motion.

Other Methodologies

There several other approaches that are applied in the detection and tracking of eye gaze.

A neural network modeling approach for finding gaze has been investi-

gated by Xu et al [162]. First, the eye appearance (relative positions of the pupil, cornea, and light reflection inside the eye socket) is modeled implicitly so that the system can effectively learn the gaze direction. Then the gaze tracker finds the focus of a subject's attention on any object on a screen. The following steps accomplish the gaze tracking procedure. 1). Eye image segmentation: to detect the small darkest region in the pupil of the eye and to segment the proper eye image. A fixed search window is used with iterative threshold, dilation and erosion methods to locate the blobs (representative objects that are like the target). In order to identify the pupil those blobs are then filtered through certain heuristics such as the number of pixels in blob, the position and value of the single darkest pixel in a blob, the ratio of blob's height of it width, and the motion constraint that the eye movement is smooth and relatively small within two adjacent sampling fames. 2). Histogram normalization: for a neural network to find features inherent in the image and to learn to associate these features and their distributions with the correct gaze points on the screen, the segmented eye image needs to be normalized. 3). Neural network modeling: the input retina units receive the normalized activation pattern of an eye image. The hidden units are divided into two groups, and connected to the corresponding two groups which encode the horizontal and vertical position of a gaze point respectively, based on Gaussian coding of output activations. 4). Training data collection: the user is prompted to look at and follow a blob cursor which moves across the screen while a camera grabs this head image sequence which, together with the grid position of the cursor, forms a training example. During this process the user is asked to satisfy certain constraints. 5). Training of the neural network: a modified back-propagation algorithm is exploited to train the neural network and an evaluation criterion called average grid deviation is introduced for stopping purpose. A two phase strategy that is used to train the neural network can lead to rapid reduction in training error which finally settles down to a table status allowing for no further overfitting of the neural network. This system works in natural office lighting environment and in real time (20 Hz) with the accuracy of 1.5° or $\pm 12mm$ if the user sits at a distance of about 22 to 25 inches away from the screen. One common video camera is mounted on the right hand side of the display screen and no other additional devices are needed.

The neural network method is often integrated with model based gaze estimation because the model based gaze tracker is usually slow and sometimes

fails to track the facial features due to rapid head movement. Stiefelhagen et al have proposed a hybrid approach to tracking interactions between people during a discussion or meeting situation [146]. A neural network is combined with a model based gaze tracker for this process in which network serves as two functions: 1).coarsely detecting gaze direction (i.e. the determination of where it is left look, right look, or down look); and 2). more precise gaze determination. With the hybrid approach the failure rate of the new gaze tracker has been reduced almost 50% compared to the model based gaze tracking.

Rikert and Jones have presented a preliminary work on gaze estimation by using morphable models [125]. Here we first introduce the ideas of morphable models and then discuss how to calculate the gaze direction by using the models. Jones and Poggio described *a multidimensional morphable model* — a flexible model for representing images of the called *a priori* objects (for example, faces), and proposed an algorithm for matching the model to a novel image and performing image analysis . The morphable model is learned from example images of objects and can be matched to a novel image by using an effective stochastic gradient descent algorithm to find the parameters that minimize the error between the image generated by the model and the novel image. An image is presented by associating it with a shape vector and a texture vector. The shape vector of an example image associates to each pixel in the reference image the coordinates of the corresponding point in the example image. The texture vector contains for each pixel in the reference image the color or gray level value for the corresponding pixel in the example image. For specifications on how to model classes of objects and how to match the model, [100] provides detailed description. The method of model-based gaze estimation involves the following steps. 1). Face detection: the face is located within the input image and a rectangular region bounding the eyes and upper bridge of the nose is extracted for matching. 2). The morphable model is built from prototype faces which span the space of head orientations, iris locations and facial appearance. 3). The model is matched to the input image so the screen position (where a person is looking) can be estimated by the parameters of the model. The head orientation and the position of the pupils in the eye sockets determine the gaze direction. Several simplifying constraints are imposed here for the initial investigation.The Stuttgart Neural Network Simulator is used in this process to generate a neural net and train it on the data. The difficulty in the model-based gaze estimation is that

a better matching algorithm is needed for matching the input image well because bad matches mean very large errors in the input to the neural net which leads to large errors in its output.

The physiological geometry of the eyes can help gaze determination. Wang and Sung proposed an approach for accurately measuring the eye gaze from images of irises [151]. Eye iris contours are modeled as two planar circles and the ellipses of their projections onto a retinal plane are estimated. And the gaze can be determined by using this circle-ellipse correspondent: the equation of the ellipse can be obtained from least square fitting of quadratic curve using the detected iris edges and with two irises it will be shown that the two respective equations lead to a unique solution of the eye gaze. There is a geometric constraint (prior knowledge) for the solution the gaze direction: the normal directions of the left and right iris boundaries are parallel to each other irrespective of eyeball rotations and head movement, which is applicable for human-machine interaction (where observed target is at a distance of more than half meter away). In this approach two cameras are used, one zoom-in camera for high-resolution iris images and another camera for determination of the head pose. Given the image of a circle in 3D space and the corresponding ellipse in the image, its pose relative to the camera frame can be analytically solved. A degenerate case, where the two iris contours are symmetrical about the plane set up by vertical and optical axis of the camera and in such a case it is impossible to distinguish from the ellipses whether the user is looking upward or downward, can be prevented in application by putting the camera slightly skewed to the face. In the proposed algorithm a simple eye model is defied so that relationship between gaze direction and parameters of the eye can be set up.

Kalman filtering is often used in gaze detection and tracking process. It is not directly related with the calculation of gaze orientation but involves in the pupil tracking which is the first step for gaze tracking. A Kalman filter is a set of recursive algorithms which estimate the position and uncertainty of moving objects in the next frame. Ji and Yang discussed the Kalman filtering in pupil tracking [98]. Kalman-filtering-based pupil tracking can be described as follows. The image sequence is sampled and the state vector of a pupil at each sampled moment is characterized as its position and velocity. According to the theory of Kalman filtering, the state vector at the next time frame linearly relates to the current state by two factor matrices: one is state transition, the other system perturbation. If pupil movement between

two consecutive frames is small enough, the state transition can be parameterized with a very simple matrix. A fast feature extractor is introduced for estimation of position of pupil. Estimated state and covariance matrix are defined according to the characterization of the uncertainties associated with the prior and posterior state estimates. The Kalman filtering algorithm for pupil tracking is realized in two steps: state prediction and state updating, which are described in detail in [98].

HMMs(Hidden Markov Models) have been used in the areas of speech and gesture analyses and recently lend a hand to the gaze control and face recognition. Their application in gaze studies are deeply related with two concepts: fixation and saccade of the eye. Fixations are brief epochs(about 300ms) in which the fovea is directed at a point of interest, gaze position remains relatively still, and pattern information is acquired from the scene. Saccades are ballistic, very fast sweeps(speed of about 900°/sec) of gaze position across the scene. About three times each second human visual system reorients the fixation point around the viewed scene through saccadic eye movements through which visual information acquisition is severely limited [84]. Henderson et al have integrated the study of gaze control for face learning and recognition across humans and artificial agents within a HMM. Understanding human's ability to control the gaze direction in order to properly orient the fovea to the interested regions is critical in the case of face perception. So the answers to following questions are essential: how potential fixation targets are selected by human; what stimulus factors determine which target is fixated; why choose these fixation sites over others; to what degree the certain sequence of fixation is important in face learning and recognition? In their studies, ten regions of interest are defined for each face based on the human fixation patterns and a left to right HMM corresponding to these regions is built by using input from a foveated vision system. An observation sequence consisting of 30 observation vectors is produced for each image and the system fixates three times in each of the ten regions. The HMM will be built when all of its parameters have been learned (different people have different parameters with same states). Each HMM is built based on 5 images for each person (class) and another image is used for testing of recognition performance. The simulated fovea is a software-defined square patch with center at the expected fixation point and ten regions of each face are foveated. The experiment consists of study phase and recognition phase and results in promising conclusions for the studies on human cognitive process

and machine vision. For example, human observers, when allowed view faces freely, make clear choices about which features to orient to and which to ignore; the selection of a fixation site during learning is driven, in part, by the specific feature that is currently under fixation; it seems much likely that feature processing during learning influences the selection during recognition; and the facial features selected for fixation during recognition of an upright face are very similar to those selected for fixation during recognition of an inverted face [84].

Salvucci also mentioned HMM's application in description of *EyeTracer* — an interactive environment for manipulating, viewing, and analyzing eye-movement protocols [128]. The HMM algorithms, used to determine the most probable interpretation of a protocol given a probalistic model of behavior in fixation tracing, can alleviate typical noise and and variability in eye-movement. To trace a sequence of identified fixation, a tracer HMM is constructed that embodies the given process model. This HMM in fact contains many smaller fixation HMMs that, in turn, represent fixations on the predicted target areas. Each of these sub-HMMs incorporates probability distributions for the location of observed fixations, centered around the center of the predicted area. The fixation HMMs are then combined into a tracer HMM that encapsulates the predictions of the model grammar by which, fixation tracing interprets the observed fixation sequence using standard HMM decoding process, thus alleviating noise and variability. With other algorithms, the HMM identification method produces very good results in fixation identification (translating a raw eye-movement protocol to a sequence of fixations and saccades) and in fixation tracing (interpreting protocols by mapping observed protocols to the sequential predictions of a cognitive process model).

PCA(Principle Component Analysis) technique can be used to in gaze tracking system. Talmi and Liu introduced a video-based, contact-free measurement system that allows combined tracking of the subject's eye positions and the gaze direction in real time [148]. Firstly, the general characteristics of the human eye are learned from various people with PCA approach and stored as refernce eye patterns. Next, the current input image of one camera is analyzed and compared with the eye patterns so that the locations of the eyes. Then the determined eye regions are searched in another camera image and thus the 3D positions of both eyes can be calculated. Finally, information us used to control the pan-tilt eye camera (the third camera) with

IR-LED light source so that the gaze direction with compensation of head movement can be determined. Here PCA technique is applied to model the human eye with the aim to describe and represent the general characteristics of human eyes with only very few dimensions, fulfilling the first part of the gaze tracking processes proposed above. In the following we just give a brief discussion how it works (specific descriptions can be found at [148]). The luminance description of the eye is transformed in a new coordinate system which has the specific property that the mean-square error introduced by truncating the basic vectors of the coordinate system is a minimum. From a set of typical pictures of eyes (eye regions), the mean of the eye regions and the difference of each eye region from the mean can be calculated respectively. The principle components of the eye space are eigenvectors of a covariance matrix and some eigenvectors (eigeneyes), representing the major characteristics of human eyes, can be selected from a set of the eigenvectors calculated by Karhunen-Loeve-Transformation. A new input image is projected into the eigeneyes space so that the similarity is measured in order to see if the new image can be classified as an eye region.

There are many other methods that have been incorporated into the gaze detection and tracking system. For example, color model based on the distribution of skin colors and the localization of eyes as a complement [39] [48]; iterative thresholding for searching and tracking eyes/pupils [48] [79]; a dual-state (eye open/blinking)for tracking eye features [24] [79] [81] [109].

0.2.3 Applications

Gaze trackers have been becoming available in technique and price ranges that allow them to be used as a useful interface component. It has been demonstrated that gaze/eye trackers are efficient in some special circumstances. But to design a human-computer interface with eye movement, several interface design issues must be considered. First, just as Jacob said, "the application of eye movement input is to make wise and effective use of eye movements, ideally in a non-command-based style"[18]. Eye movements must be interpreted carefully to avoid annoying the unwanted responses to user's actions because of the nature of non-consciousness in eye input. Next, more specifically, eye movement input is very faster than other input media and it is extremely simple to operate: any normal users can use their eyes to look at any object rather quickly. What is more, a user view s a single object

with a sequence of several (distinct, individual) fixations, all in the general area of the object. Finally when eye tracker fails to obtain an adequate video image of the eye for one or more frames, instability may occur into the output of the eye tracker. These issues have remained a challenge in the application of eye movements as an HCI input media. In the following discussions of the eye/gaze application, we have not mentioned how these problems have been solved to a larger or less degree. We only introduce the to what area the eye/gaze tracking has been applied.

Comparison of eye gaze interaction technique for object selection with the traditional method of selecting with a mouse was made by Sibert et al. Their conclusion is that "our eye gaze technique is measurably faster", that "eye gaze interaction has additional harder-to-quantify benefits", and that "our interaction technique design and implementation preserves the inherent speed advantage of the eye over the hand" [142].

Based on the studies of eye movement (such as saccades, fixations, and blinks), an advance application is to develop eye-aware software that can adapt in real-time to changes in a user's natural eye-movement behaviors and intentions. Edwards introduced such kind of development tool (Eye Interpretation Engine). During the process of its development, they discovered two important features of eye-movement patterns: revisit (a fixation that goes back to the location of a recent previous fixation) and significant fixation (a fixation that lasts longer than a variable threshold). Both patterns complement the recognition of saccades, fixations, and blinks, and can make easier the recognition of high-level patterns in user's natural eye-movement. Identification of the eye-movement patterns involves three behaviors. 1). Knowledgeable movement: default behavior that becomes active after each significant fixation until another behavior is recognized. 2). Searching: it is defined by the presence of more than one large saccades since the last significant fixation, or the presence of a series of consecutive small saccades that occur since the last significant fixation and that collectively cover a distance greater than or equal to a large saccade. And 3). Prolonged searching: defined as 10 or more saccades since the significant fixation. One part of this tool is an on-screen keyboard system that has keys that can be activated by looking at them. This keyboard system can help improve the interaction between user and computer, that is, the computer adapts to the human user (the eye-activated keyboard is fast and effective enough for experienced users and easy enough at same time for first time users). Another part of

the tool is to develop a toolkit that has the power of being able to recognize the different behaviors so a software written with the tool can adapt to the changing needs and expectations of the user [77].

In the following we introduce several HCI systems that use gaze as one of their modalities.

ERICA (Eye-gaze Response Interface Computer) is computer system that tracks eye movement, with which individuals can perform all actions of mouse and keyboard with their eyes in Window platform. ERICA is especially useful for the disabled. Lankford introduced this system in [105] [106]. Two subsystem are included in ERICA: *eye-mouse* and *eye-typing*. Gaze clicking is realized for the eye-mouse function by counting the fixation time at a point on the screen. Colorful shapes serve as a visual cue to the users that if they keep fixating at that point, they will perform a mouse control action at the point. For example, red rectangle, blue circle, green rectangle have different meanings depending on the amount of the gaze fixation time and on the actions the user has taken. And zooming techniques have been used to execute a mouse action at a desired location more accurately and reliably. Only an object the user wishes to interact with needs to have an increased size. Also a variety of smoothing algorithms have been used to remove jitter in the mouse cursor. The ERICA's eye-typing (called Visual Keyboard) grants the user the means to perform text entry functions. Prolonged fixation on a key on screen causes that key's action to be performed. Dwell time activation in eye-typing can be as little as 0.5 seconds. A dictionary will be shown at the bottom of the screen narrows the word selection choice to match the characters typed by the user. The keyboard's dictionary can update in response to the user's typing and mouse actions. Three types of keys are included in the Visual Keyboard. And as the user shifts from application to application, the keyboard's layout and dictionary change to suit the particular application. The user may also use eyes to reconfigure the keyboard layout. Isokoski proposed another text input method with eye trackers of using off-screen targets [97]. A limited number of targets outside the computer display are used to input a limited number of different tokens into a computer. Several coding methods have been introduced for the function of how to combine those limited number of tokens and how to interpret the combination into the normal keys.

Duchowski et al described the application of eye/gaze tracking in a VR (Virtual Reality) system for aircraft inspection training [75]. The user's

scanpaths (gaze locations), which are recorded through tracking the gaze direction, are calculated as gaze/polygon intersections, enabling comparison of fixated points with stored locations of artificially generated defects located in the environment interior. In VR,the 3D gaze locations serves as a post-immersion diagnostic indicator of the user's overt focus of attention. The collection of scanpaths over the course of immersion is used as a diagnostic tool for post-immersive reasoning about the user's actions in the environment. The effects of training, via comparison of the performance of experts to that of novices, can thus be gauged. A binocular eye tracker embedded in the system head mounted display is integrated into the VR inspection system to realize the evaluation of the training process.

There are other gaze tracking applications. In realistic and comfortable 3D representation, tracking of subject's eye position and gaze direction is integrated into an autostereoscopic display that supports the natural link between accommodation and convergence of human vision, reproduces the limited depth of focus of human vision and shows comfortable, hologram-like scenes with motion parallax to subject [148]. During face-to-face communication or in a meeting environment, finding out at whom or where the speaker is looking is helpful to understand whom the speaker is talking to or what he/she is referring to, to give cues of people interest and attention, and indicate interpersonal cues by automatically monitoring the gaze of participants in a meeting room [146]. Face recognition is a well-studied, but not completely solved problem. The study of gaze control with human can be integrated into the investigation of face perception and recognition [84]. Facial caricaturing introduces the gaze direction and distribution into its application [116]. Gaze detection and tracking has been exploited to the studies of driver's vigilance and fatigue [98] [52]. A system is introduced also for classifying the focus of attention of a car driver [120]. Because drivers are usually not aware of their eye movements which can provide the visual information acquisition strategies applied by drivers, analysis of driver eye movement data gives a far more quantitative method for evaluation of traffic control devices than subjective evaluations. An instrumented vehicle was designed for recording and analyzing the eye movement including gaze tracking [131]. A gaze-based control system in Boeing flight deck simulator was set up to simplify operations in commercial flight decks by reducing the number of dedicated mechanical switches and control panels [131].

0.3 Current Research in Head Gesture and Head Pose

The development of information technologies, especially that of computer science and engineering, is rapidly embedded into our environment which has imposed needs for new types of human-computer interaction, the interfaces that natural and easy to use. Human gestures such as head movement, hand posture and motion are the most powerful modalities for HCIs. Because head/face pose has been involved in the study area of head gesture, we outline the current development of studies in both areas in this section.

0.3.1 Head Pose and Eye Gaze

The features of face are often used to construct 3D face model from which head pose and gaze direction can be calculated. Newman et al proposed a real-time system in which two cameras are arranged to track the 3D head pose and gaze [118]. The tracking system consists of five parts. 1). 3D facial model creation: up to 32 features of face have been selected that help result in better tracking, for example, corners of the eye and mouth, nose region, ears and eyebrows, and etc. The 3D model is created through taking snapshot of the head; identifying the features in the stereo images; identifying the corners of the eyes, the features of the side of the head; and taking snapshots of the head turned 45° to the left and right. 2). Face acquisition by constructing a face template and matching it in the tracking procedure. 3). 3D face tracking is made by using the templates and mapping the model to the measurements. 4). Gaze point estimation by modeling the eyes as spheres and extracting the pupil centers and iris boundaries. 5). visualization: animation of 3D surface model of and face and a cone representing the gaze direction. In this system a mannequin head is used mounted on a pan-tilt device to demonstrate the accuracy of the head pose estimation, which illustrates very promising results.

In a similar system, Heizmann and Zelinsky used a three layered system for estimating 3D facial pose and gaze point [91]. The lowest level correlates to the hardware. Its measured feature positions are forwarded to the 2D model level (middle level) which takes into account geometric constraints in the image plane and the correlation distortion. The 2D image positions of the features are transferred to the 3D model (high) level, where 3D pose of the

head can be determined and used for gaze point. In order to effectively calculate 3D pose and gaze, Matsumoto and Zelinsky designed a system which has the following techniques: 3D vision hardware with a field multiplexing device, image processing board with normalized correlation capability, 3D facial feature model, and 3D eye model [113]. Because the 3D coordinates of the facial features can be directly measured, the algorithm for computing 3D head pose (and also for 3D gaze) has been greatly simplified.

In Feature-based tracking system, the 3D eye gaze direction can be calculated even with head rotation and using a monocular camera. Furthermore such system can automatically initialize the feature tracking and recover from total tracking failures which may occur when a user becomes occluded or temporarily leaves the image.

Head orientation and gaze direction are important indicators for determining driver's fatigue and tiredness and thus will be significant factors in the development of automatic safety mechanisms, in which specific facial features are not so commonly used as in other research areas. Pappu and Beardsley presented an approach to classifying the focus of attention of a car driver [120]. Head pose is a good, although insufficient, indicator of driver's attention. Four components to this method are: 1). the driver's head is modeled with an ellipsoid; 2). this ellipsoid is used to generate an array of synthetic views for a range of head motion including rotation; 3). a target image of the driver is matched against the synthetic views; and 4). information about head pose is accumulated over time in a pose-space histogram. For any acquired image of the driver, its best-matching synthetic view can be found and used to index the corresponding location in the histogram so that the target image can then be classified according to the labeling of the nearest peak in the histogram, achieving the classification of the driver's focus of attention.Because no explicit Euclidean measurements are needed in this approach, there is no requirement for such a priori knowledge as car's interior geometry, the camera calibration, or driver's exact location. The system runs on as SGI workstation at about 6Hz.

Ji and Yang also proposed model-based approach to face pose estimation in a driver-monitoring system [98]. Their method is based on the following observations: 1). the inter-pupil distance decreases as the face rotates away from the frontal orientation; 2). the ratio between the average intensity of two pupils either increases to over one or decreases to less than one as face rotates away or rotates up/down; 3). the shapes of two pupils become more

elliptical as the face rotates away or rotates up/down; and 4). the size of the pupils also decrease as the face rotates away or rotates up/down. A face pose estimation algorithm (called pupil feature space) can thus be developed by exploiting the relationship between face orientation and the pupil parameters: inter-pupil distance, and sizes, intensities, and ellipse ratios of both pupils. The head orientation is quantized into 7 angles from -45° to 45° and the correctness is above 94.67%.

0.3.2 Head Pose in Single Image

Head pose can be calculated from one single image. Chen et al presented an approach to estimate the 3D head pose in a single image without exploiting facial features [85]. Only the information about the skin and hair regions of head is required for the estimation. At first, face is detected and the skin regions and hair regions of face are extracted by using a perceptually uniform color space to describe the color information of images and estimating the skin color likeness and the hair color likeness. Then the area, center and axis of least inertia of both the skin region and the hair region are calculated by introducing the *densities* of skin and hair regions, thus setting up a set of equations for the area, center and axis of skin and hair regions, and then solving for the results. Finally, the pose of a face in an image is the result of the combination of three rotation elements in a 3D coordinate system. The results in the above step are used to calculate the three elements in simple head rotation and complex head rotation. Because this method concerns with only global information such as area, center about the skin and hair regions, the extraction of which is stable and not sensitive to local changes of facial features, it is relabel and easy to use. But the speed is relatively slow.

A different method of determining head pose in one image was proposed by Shimizu et al [83]. The basic idea is to use a generic 3D model of head for the variation in shape and facial features and project this 3D generic model onto image plane, and then the contours of eyes, lips and eyebrows are defined as edge curves for the 3D model. After the correspondences between the edge curves on the model and the edges in the image (the distance between curves is set up in the process) are established, the head pose can be estimated by using ICC (Iterative Closest Curve) method. With a reasonable initial guess, ICC minimizes the distance between the curves on the model and the corresponding curves in the image. For more precise pose calculation,

variable edges, which are pieces of occluding contours of a head, is used. This approach does not require the internal parameters of a camera, but can deal with the shape difference between individuals and get accurate head pose because of the use of 3D generic model.

0.3.3 Several Algorithms in the Estimation of Head Pose

There are several algorithms that can effectively be used to estimate the head pose, for example Gabor wavelet networks, EM algorithm. Kruger et al introduced an efficient head pose estimation method with GWN(Gabor Wavelet Networks) [104]. GWNs represent an object as a linear combination of Gabor wavelets and the parameters of Gabor functions can be optimized to illustrate the specific local image structure. One reason for using GWNs in pose estimation is that they are by nature invariant to some degree to affine deformations and homogeneous illumination changes. Other reasons include: 1). Gabor wavelets being recognized to be good feature detectors in that an optimized wavelet has the exact position and orientation of a local image feature, thus leading to a higher level of abstraction in representing an object if Gabor filters are used as a model for local object primitives; 2). the weights of each of the Gabor wavelet being directly related to their filter responses and to the underlying local image structure; and 3). by simply varying the number of used wavelets, the precision of the representation having a wide range of variations, from a coarse representation to an almost photo-realistic one. An experiment with a doll' s head showed that the minimal mean pan/tilt error was 0.19° or a GWN with 52 wavelets, 0.29° with 16 wavelets, and that the maximal error were 0.46° for 52 wavelets and 0.81° for 16 wavelets. And speed could be expected to more than 5 fps for 52 wavelets and 10 fps for 16 wavelets.

Choi et al proposed a method for estimating 3D facial pose by using EM algorithm [71]. The underpinning philosophy of this approach is that the EM (Expectation-maximization) algorithm can be exploited in the process of learning the 3D pose parameters subject to constraints provided by the location of the bilateral symmetry axis of the face and the orientation of the line connecting the two eyes and that EM algorithm has been successfully adopted as a registration engine in matching line-templates, shape templates

and 3D perspective models. For the first step in the approach, a generic 3D template of the facial features is constructed and projected onto the 2D feature locations. The left and right eyes, the lips and the chin are thought to be coplanar, the tip of the nose resides at some height above the plane, and the planar features are symmetric about the axis defined by the center-points of the lip and the chin. The projection of the template onto the locations of the 2D facial feature points has six degrees of freedom: two translation parameters on the 2D image plane, the overall isotropic model scale, and the three Euler angles that define the 3D rotation of the model points. Three degrees of freedom (two template translation parameters and an Euler rotation) can be removed by centering and aligning the template at a fixed point on the bilateral facial symmetry axis. Next is the expectation step, which involves estimating a mixture distribution using current parameter values. The EM algorithm provides an interactive framework for computing the a posteriori matching probabilities using Gaussian mixtures defined over a set of transformation parameters. Finally comes the maximization step that involves computing new parameter values that optimize the expected value of the weighted data likelihood. Experiments for evaluating this method (based on both contrived data with known ground-truth together with some more naturalistic imagery) have shown that the algorithm can estimate facial pitch within 3 degrees if the overall rotation does not exceed 40 degrees.

The shape-from-shading algorithm can be an alternative method for the estimation of facial pose. Shape-from-shading is based on the idea that local regions in an image correspond to illuminated patches of a piecewise continuous surface whose height is represented by a function of the coordinates. The observed image intensity will vary depending on the material properties of the surface, the orientation of the surface at the coordinates, and the direction of illumination. Choi et al developed a new framework for shape-from-shading and applied it to the problem of facial pose estimation as the first step in a more ambitious program of work aimed at using the improved needle-maps for face analysis [71]. Their approach is a geometric one. The image irradiance is considered as a cone of ambiguity about the light source direction for each surface normal. A valid needle-map can be recovered by satisfying the irrdiance equation at every iteration and the task of shape-from-shading thus becomes that of iteratively improving the organization of the needle-map using curvature consistency constraints. And then the facial pose can be recovered by suing orientation histograms extracted

from the needle-maps delivered by shape-from-shading. First, a simple geometric model is constructed to represent the way in which the distribution of needle-map directions transforms under rotation of the head. Then the changes in the histogram of orientation angles are to be computed under horizontal and vertical rotations. Finally the pose is recovered by locating the rotation angles that lead to maximum similarity between the observed orientation histogram and the transformed model histogram. This method can yield head pose only within a few degrees because shape-from-shading is not mainly used for pose estimation but for face recognition.

SVMs(Support Vector Machines), based on a generic learning framework, have shown great potential for learning classification functions that can be applied to pose estimation. Ng and Gong extended SVMs to model the 2D face appearance that enables simultaneous multi-view face detection and pose estimation [119]. They have done this by introducing SRM (Structural Risk Minimization), which can determine a function which best captures the true underlying structure of the vectored-encoded data and provides a well defined quantitative measure for the capacity of a learned function to generalize over unknown test data, thus offering a guaranteed minimal bound on the test error. A multi-view face model is constructed by using SVMs and the model is then applied to pose estimation across the view sphere. The boundaries of the face pose distribution is defined by support vectors, which are localized with regard to the pose sphere and can be effectively used to perform pose estimation by using nearest-neighbor matching. This method can not applied in real-time tracking, but is very promising if using the non-linear mapping learned by the SVM classifier instead of the current simple nearest-neighbor matching in retrieving pose information from support vectors.

Head orientation can also be computed by using an ellipsoidal model of head. Wu and Toyama presented an algorithm for estimation of head orientation from cropped images of a subject's head from any viewpoint [161]. An ellipsoidal model is created of points, each of which maintains probability density functions of local image features of the head based on training images. Edge-density features are extracted by using a Gaussian at a coarse scale and rotation-invariant Gabor templates at 4 scales, and then projected onto the model to collect data for each point on the model. Each model point learns a probability density function from the training observations. During pose estimation,maximum a posteriori estimation is implemented from input images by using different priors tailored for the

cases of global pose estimation and pose tracking. This method has the advantage of being insensitive to common variations in facial appearance, being illumination-insensitive, being able to handle side and back views, and working under wide range of image scales and resolutions. But the accuracy in pose estimation is low and the speed also slow.

In the problem of understanding pose discrimination in similarity space, Sherrah et al have found that PCA(Principal Component Analysis) is "an appropriate representation for pose similarity prototypes because it suppresses identity variations while maintaining sensitivity to pose" [140]. Two issues are to be considered: 1). for a given pose, what transformation of the images is optimal to exaggerate differences in pose while suppressing differences in identity, and 2). what is the minimum angular separation that can be resolved using similarity-based methods? A criterion is defined so as to quantify the goodness of a given transformation method for pose prototypes. Three experiments have been tested on the criterion: 1). Gabor filters are examined as a method for enhancing pose differences at each pose angle; 2). PCA is used to represent prototypes and its identity-invariant properties are examined; and 3). the criterion is used to determine the angular resolution at which neighboring poses can be resolved. The results are: orientation-selective Gabor filters enhance differences; different filter orientations are optimal at different poses; and PCA has the advantage of proving invariance to identity while accurately describing pose changes and of being understandable through visualization and more computationally efficient. The lowest angular separation in pose difference can reach to lower than 10°.

A new concept of *perceptual fusion* has been applied in tracking head pose by Sherrah and Gong [136]. The philosophy of the concept is that multiple sensory modules are integrated to arrive at a single perceptory output. A conditional density propagation algorithm *CONDENSATION* is adopted for the tracking task. CONDENSATION is a particle filtering method which models an arbitrary state distribution by maintaining a population of state samples and their likelihoods. It is more generic and flexible than Kalman filter because of its propagation of arbitrary density models and the state samples capable of multiple hypotheses for the current state of the system. Experiments have shown that the system can track the head tilt and yaw angles and recover from momentary loss of lock on face position and head pose. Without the fusion the tracker will wander away to incorrect poses or non-faces.

0.3.4 Face Features and Pose

Many facial pose estimation approaches either are only a part of and thus integrated in the larger systems of face recognition or are examined with the facial feature analysis. Those features play very important role in the calculation of face pose.

Elagin et al presented an automatic module that can determine the pose of a human face from a digitized portrait-style image on the base of Bunch Graph Matching [78]. By matching series of bunch graphs, each of which represents the pose with a certain degree of rotation to the input image containing mug-shot of a human head, the pose of a bunch graph that best matches the input image is returned as the pose the face in the image. At first, the face knowledge is employed in the form of a graph, the nodes of which are labeled with image information referring to landmarks or local areas on the face like the pupil, the tip of nose, and the edges of which are labeled with distance vectors between the nodes. Then a face can be detected by dragging a bunch graph representing general face knowledge across the whole image checking similarity at each point. Finally, face pose is estimated by using a set of bunch graphs that capture fine details of the image so that a decision about the pose can be made. In this process the concept of similarity threshold is exploited. The system runs at speed of close to real-time performance and pose estimation success rate was about A98.5% for a set of 210 faces rotated in various degrees and directions.

A similar method was proposed by Kruger et al for estimation pose [103]. Their system, an extension of a face recognition system to pose estimation, can automatically determine the position, size, and pose of head. Face is represented with bunch graphs as the above. A *total similarity* is defined between a grid on a certain position on an image and a bunch graph representing a certain pose with a certain size. Then a bunch graph is adapted to an image by Elastic Graph Matching for the estimation of head pose. Pictures of head with different view orientations show that this algorithm can simultaneously solve head finding, scale normalization and pose estimation.

Borovikov described a different method for head pose estimation by using facial feature location [67]. A set of *crucial facial features*, which mean some of the facial features that can be a major clue to the head pose recognition, are located within the head silhouette boundaries and then are used to recover the head orientation in 3D space. The process includes head sil-

houette estimation, crucial facial features location, best constellation search, and estimation of the head pose. The method works with both color and gray-level images from quasi-frontal of a human head under variable lighting conditions.

Pose invariant face recognition is a difficult problem because the pose of face can greatly affect how the face appears in an image. If the facial orientations are known the process of face recognition can become relatively less difficult. Lee and Rnaganath once described a face recognition system based on a recognition-by-synthesis approach [108]. System first estimates the pose of the unknown face, then synthesizes the face images of known subjects in the same pose as the unknown face for the recognition. The pose of the face is computed by matching an image of an unknown face to a 3D deformable face model which can encode shape as well texture. This model is a composite of three sub-models: 1). edge model that defines the outlines of the face as well as various facial features such as eyebrows, nose, and etc, 2). color model that that identifies facial regions of low intensity, high intensity or fairly homogeneous lip color, and 3). wire frame model that approximates the 3D structure of the face and can be used to synthesize face images. The 3D face model is matched to face images by using complex deformation and geometric transformation, yielding the facial pose. The accuracy in pose estimation by this system can reach up to with 5°. And as a face recognition system it can handle large pose differences where appearance changes significantly and occlusions occur in parts of the face.

0.3.5 Head Gesture Detection and Tracking

Yachi et al proposed a method for detecting and tracking a human head in real time from image sequence [163]. A fixed-viewpoint pan-tilt-zoom camera is used to acquire image sequence and the variations in the head appearance can be eliminated due to camera rotations with respect to the viewpoint. A variety of contour models of head appearances are related are created to be related with the camera parameters so that the model can be adaptively selected to deal with the variations in the head appearance due to human activities. The model parameters obtained by detecting the head in the previous image are used to estimate those to be fitted in the current image. As a result human head can be robustly detected against its appearance variations and be tracked in near real time.

A different method for head tracking, via robust registration in texture map images, was presented by Cascia et al [70]. A texture mapped 3D surface model is used for head, and tracking is formulated in terms of color image registration in the texture map of a 3D surface model. Model parameters are recursively updated via image mosaicking in the texture map as the head pose varies. The output, the dynamic texture map, is the 3D head parameters and a 2D dynamic texture map image which can provide a stabilized view of the face that can be used for facial expression recognition and other applications in tracking. A robust minimization procedure is exploited in the tracking system so that it is almost insensitive to eye blinking and robust to occlusions, wrinkles, shadows, and specular highlights. But the initial positioning of the model must be done by hand and the system lacks a backup technique to recover when the track is lost.

Schodl et al also exploited 3D texture model in tracking head but implemented in different ways [27] [28] [5]. The implementation of the 3D texture model-based head tracking system is realized by using Gaussian pyramids to deal with large head motions and OpenGL to perform visible surface determination and rendering of the the textured model. During tracking, six translation and rotation parameters are found , by mapping the derivative of the error with respect to the parameters to intensity gradients in the image, in order to register the rendered images of the textured model with the video images. Sequential version of this method is implemented by using nonlinear conjugate gradient with variable step size. The minimization of parameters in low resolution is used for finding the final parameters in the higher resolution level. The model is rendered via OpenGL for each gradient computation in the conjugate gradient algorithm. The system has also been implemented in adaptive parallelization on an eight node workstation cluster, by performing multiple error and gradient evaluations simultaneously on each node, in order to get it closer to real-time performance. This method has similar stability as sequential conjugate gradient does and computation is much faster at about .25 s/frame.

0.3.6 Miscellaneous Head Gesture Studies

In this part we focus on miscellaneous topics on head gesture studies, for example, features of head [82] [152] [74] [101] [87] [95], head and facial expression [72], head and hand [138] [164] [135]. Here we discuss the first part.

The studies of head and facial expression, and head and hand belong to the other following sections.

Gong et al presented a method for learning appearance models that can be used to reconΦgnize and track both 3D head pose and identities of novel subjects with continuous head movement across the view-sphere [87]. This approach has been embedded into an automatic face data acquisition system, which systematically obtains a database of face images with labeled 3D poses across a view-sphere of $\pm 90^\circ$ yaw and $\pm 30^\circ$ tilt at intervals of $a0^\circ$. This database is used to learn appearance models of unseen faces based on similarity measures to prototype faces.

A multi-model system is designed [82] for locating heads and faces in that three channels are combined: 1). shape analysis on gray-level images that determines the location of individual facial features and the head outlines, 2). color segmentation that aims to find areas of skin colors by using a clustering algorithm, and 3). motion information that is extracted from frame differences so that head outlines are determined by analyzing the shapes of areas with large vectors. Combinations of shapes produced by the three channels are evaluated with n-gram searches to yield the likely head positions and facial feature locations. The accuracy for finding head positions and facial features can reach up to 95% and 93%.

A tracking method, image-based parameterized tracking for face and face features to locate the area in which a sub-pixel parameterized shape estimation of the eye's boundary is performed, is employed for estimation 3D head orientation in a monocular image sequence [95]. Four points at eye corners and the tip of nose are involved in the the sub-pixel parameterization estimation process. The 3D head orientation is computed through employing projective invariance of the cross-rations of the eye corners and anthropometric statistics to estimate the head yaw, roll and pitch.

Weber et al presented a model learning and training'Φ method for detection of human heads from different viewing angles [152]. Only a model where objects are represented as constellations of rigid features is considered and a joint probability density function on the shape of the constellation represents variability. Distinctive features are automatically identified in the training set by using an interest operator followed by vector quantization. And the the set of model parameters (including the shape probability density function) is learned by using expectation maximization. Experiment results have shown that the performance is above 90% correct with less than $1s$ per image

to novel viewpoints and unseen faces.

An approach for detecting nodding and head-shaking in real time from a single color video stream by directly detecting and tracking a point between the eyes is proposed in[101]. Once this "between-eyes" is detected, a small area around it is copied as a template by which the tracking is done at 13 frames/sec. By analyzing the movement of the point, nodding and head-shaking can be detected.

0.4 Current Research in Hand Gestures

Human hand gestures, from the simplest actions of pointing at objects to the most complex ones that can express our feelings, are a kind of non-verbal interaction among people. In the early days of studying hand gestures, subjects were asked to wear glove-based devices, resulting in unnaturalness and obstructiveness. Vision-based approaches to hand gesture researches, by using a set of video cameras and computer vision techniques to interpret gestures, can overcome these limitations. The recognition of static hand gestures (postures) have obtained many successful results by setting up various models (by using general object recognition approaches) such as images of hands, contours, silhouettes, and 3D hand skeleton models. Yet human hand gestures are a process of natural dynamic actions and the motion of the hands conveys meaningful clues. In this section we survey the researches in both static and dynamic areas of computer-vision-based hand interfaces. We based our review on Pavlovic et al's paper [121], Wu and Huang's paper [158] [160], and included lots of other papers.

Many approaches to the visual interpretation of hand gestures have been focused on a particular aspect of gestures and there have been numerous approaches to Hand-centered HCI problem. Pavlovic et al proposed a global structure of the interpretation system to study the process of hand gesture interpretation [121]. In that system, a mathematical model of gestures is created and will be pivotal for the successful functioning of the system. Video input streams are fed into the analysis stage where model parameters are calculated from image features which are extracted according to the specific task of gesture interpretation. Then the parameters are classified and interpreted in the recognition stage in the light of the accepted model and the rules imposed by an adequate grammar which reelects both the internal

gestural command syntax and the possibility of interaction of gestures with other communication modes like speech and gaze.The output of recognition stage is the gesture description, which gives a measurement of naturalness of interpretation of gestures, including accuracy, robustness, speed, and different classes of hand movements. Some of the output can also be fed back into the analysis stage. We will discuss the methodologies and techniques in modeling, analysis, recognition of hand gestures and the applications of those techniques.

0.4.1 Hand Gesture Modeling

Different applications in the hand-centered HCI decide the different kinds of hand gesture models, some of which may be simple, other more complicated. For better describing the models, Pavlovic et al used a vector that describes the pose of the hands and their spatial position in the parameter space to represent the hand gesture by a trajectory in the parameter space over the defined interval. Furthermore, they went on to provide a general gestural taxonomy for HCI, which can influence the way parameter space and gesture interval are determined. According to the taxonomy, the most commonly used gestures in HCI context are symbols: the gestures that have a linguistic role and belong to the communicative gesture category.

Temporal Modeling

The determination of gesture interval is related to temporal(dynamic) characteristics of the hand gestures in the form of temporal segmentation of gestures from other unintentional hand movements. Pavlovic formulated a set of rules that determines the temporal segmentation of gestures: gesture interval consisting of preparation, stroke, and retraction phases; hand pose during stroke following a classifiable path in the parameter space; gestures being confined to a specified spatial volume; repetitive hand movements as a kind of gestures; and manipulative gestures having longer gesture interval lengths than communicative gestures. Because of the complexity of gestural interpretation in the allowed temporal variability of hand gestures, the work in hand gestural studies has been largely confined to the areas of hand pose.

Spatial Modeling

Two major approaches have been used in gesture modeling. One is called 3D hand model, the parameters of which are joint angles and palm position of the hand. It is based on the hand skeleton that consists of wrist bones, palm bones, and finger bones. DoF (Degree of Freedom) can be used to describe the the joint connecting characteristics. And the hand joints can also assume a certain range of angles.3D hand models offer a way of complete modeling of fall hand gestures and lack the simplicity and computational efficiency, and thus are less feasible with the other kind of model: appearance-based models.

This second group of models is based on the appearances of hands in the image and model gestures by relating the appearance of any gesture to the appearance of the set of predefined template gestures. Some of the parameters of this model are: images, image geometry parameters, image motion parameters, and fingertip position and motion. There have been mentioned in [121] [158] [160] a large variety of models that belong to this group: deformable 2D templates being the sets of points on the outline of an object that are used as interpolation nodes for the object outline approximation, 2D hand image sequences as gesture templates where each gesture is modeled by a sequence of representative image n-tuples and each element of the n-tuple corresponds to one view of the same hand, and the models with hand image property parameters such as those derived from contours and edges, image moments, image eigenvectors, and fingertip and palm positions. Others include construction of 3D deformable point distribution model of human hand [90], gesture modeling by using FSM (Finite State Machines) [93], finger tips as templates [129].

In a real-time gesture controlled interaction system for modeling, analysis, and recognition of continuous dynamic hand gestures, a spatio-temporal appearance model is proposed for modeling by hierarchically integrating multiple cues [166]. At first, flesh chrominance analysis and coarse image motion detection are combined to detect and segment hand gestures. And then parameters of the spatio-temporal appearance model are recovered by fusion of robust parameterized image motion estimation and hand shape analysis. This approach can fulfill real-time processing and high recognition rates.

0.4.2 Gesture Analysis

During hand gesture analysis, three steps are processed: 1). hand localization and segmentation, 2.) hand feature extraction, and 3). hand model parameter computation. The results from the former step are the input to the latter step, and finally the parameters (trajectory in parameter space) are estimated for gesture recognition and tracking.

The hand localization and segmentation, a process in which the hands are extracted from the image, has applied a variety of restrictions. One is the restrictions on background. For example, the segmentation task can be greatly simplified if in a uniform, distinctive background. Restrictions on user is another one: wearing special clothing. Other restrictions include those on imaging (for example, on-hand focused cameras). Directly thresholding the image can be used for extraction of the hands from the background. Color histogram analysis is applicable in less restrictive setups because of the characteristic histogram footprint of the human skin. Moving artifacts that are produced under certain restrictions can also be used to segment out the hand from other static objects.

A hand segmentation method is proposed [165] for segmenting hands of arbitrary color in a complex scene. It is a Bayes decision based statistical approach that generates a hand color model and a background color model for a given image and uses these models to classify each pixel in the image as either a hand pixel or a background pixel.

The features extracted from the images are very similar even though the extraction depends on the model of the gestures. Hand images are considered as features by themselves (for example, fingertips extracted as features in 3D models that use finger trajectories). Widely used silhouettes are one of the simplest features, extracted from local hand images in restricted background with the help of color histogram analysis. Contours, produced with different edge detection schemes, represent another group of features. Fingertip is a very commonly used feature and its locations can be used to obtain parameters of hand models. In the following we discuss some methodologies in the extraction of hand image features.

The discriminating features can be automatically selected by using multiclass, multidimensional discriminant analysis for gesture classification and the classification can be done by a recursive partition tree approximation with the hand segmentation [73].

Objects can be represented with a rich variety of image features of different types and at different scales. The image features can be zero-dimensional(junctions), one-dimensional(edges and ridges), or two-dimensional(blobs) [69]. Bretzner and Linderg presented a view-based object representation, the qualitative multiscale feature hierarchy, and showed that this representation can be used for improving the performance of a feature tracker, by defining search regions in which lost features can be detected again [69]. Their ideas are to employ qualitative relations between image structures at different scales in representation, to track all image features individually, and to use the qualitative feature relations for avoiding mismatches and resolving ambiguous matches. Laptev and Lindeberg also used a hierarchy of multi-scale image features combined with particle filtering for tracking of multi-state hand models [107]. The likelihood of hierarchical, parameterizedmodels (containing different types of image features at multi-scales) is constructed on its maximization over different models and their parameters, is evaluated, and integrated with the framework of particle filtering in the the simultaneous tracking and recognition.

In a gesture recognition system for visually mediated interaction [137], six gestures (point left, point right, wave high, wave low, go away, and come here) are collected for extracting gesture features, each of which forms a spatio-temporal trajectory. Each gesture gives a generic model which can be used for recognition of novel gestures. A condesensation is then used to track the gesture and current time frame for recognition.

Yang and Ahuja described an algorithm to extract the motion trajectories (associated with ASL) of salient features such as human palms from an image sequence [164]. The motion segmentation of the image sequence is generated by using a multiscale segmentation of the frames and attributed graph matching of regions across frames, thus producing regions correspondences and their affine transformations. Then, skin regions are determined with colors of the moving regions and the palm regions are identified based on the shape and size the skin regions in motion. Lastly, the region's motion trajectory is constructed by concatenating affine transformations that define a region's motion between successive frames.

Parameter computation is the last step in hand gesture analysis. The type of computation depends on both the model parameters and the features. There are many methods used for the computation, depending 3D models. 3D hand model-based models often employ successive approxima-

tion methods for their parameter computation, the basic idea of which is to vary model parameters until the features extracted from the model match the ones obtained from the data images. The matching begins with palm and ends with fingers. A lot of features have been used for the hand parameter computation: hand silhouettes, fingertip locations with characteristic points on the palm, and contours and edges. In deformable 2D template-based models, model position, orientation and principal components are adjusted in successive steps until a satisfactory match between the model and the image is achieved. Direct mappings are often used between the feature and the parameter spaces. Some mappings are explicitly defined and others employ interpolation on the feature-parameter correspondence tables.

0.4.3 Hand Gesture Recognition

Gesture recognition is the process where the trajectory in the parameter space is classified as a member of some meaningful subset of the parameter space. There are two major topics connected with the recognition process: the optimal partitioning of the time-model parameter space and the implementation of the recognition procedure. The implementation of the recognition procedure is of a computational issue and of a more practical nature. The more complex the model is, the broader class of gestures it is applied to, and thus the more recognition process time. So most of the 3D model-based gesture models can not be applied in real-time performance because of their more than ten parameters in description. The appearance-based models, usually restricted to a narrow subclass of HCI applications, are computationally affordable to implement in real-time applications. We discuss the first problem in the following.

An optimal partitioning of the hand gesture parameter space, largely influenced by the kind of application the HCI system is intended for, should act in the way that it produces a single class in the parameter space corresponding to each allowed gesture that minimally intersect with any other gesture class. The complete gesture model provides these portioning rules to produce a set of distinctive classes associated with each natural gesture. 3D hand model-based gesture models, because of their close to the complete gesture model, give the possibility of general gesture recognition, but have difficulty in practice due to the computational complexity. The appearance-based models, on the other hand, exhibit satisfactory recognition performance un-

der the two assumptions: 1). these models are limited within a particular taxonomical group of gestures; and 2). they disregard dynamic properties of gestures and analyze only static postures. There are several methods in performing the actual partitioning of gesture parameter space.

The appearance-based learning approaches are promising, although they often suffer from the insufficiency of labeled training data. Wu and Huang described a method to alleviate this difficulty in virtual environment application [159]. This is the Discriminant-EM, combining supervised and unsupervised learning paradigms and adding a large unlabeled set, in the case of handling small labeled training set for recognizing a set of predefined gesture commands and detecting hands.

Wilson and Bobick presented an HMM-extended method for the representation, recognition and interpretation of parameterized gesture [155]. The parameterized gesture means the gestures that exhibit a systematic spatial variation. The standard HMM method of gesture is extended by including a global parametric variation in the output probabilities of the HMM states. An EM method is formulated, by using a linear model of dependence, for training the parametric HMM. A similar EM algorithm simultaneously maximizes the output likelihood of the PHMM for the given sequence and estimates the quantifying parameters. Finally, the PHMM is extended to handle arbitrary smooth (non-linear) dependencies and a generalized EM algorithm is employed for both training and the simultaneous recognition of the gesture and estimation of the value of the parameter. This nonlinear approach permits the natural spherical coordinate parameterization of pointing direction. The similar method can be found at [156] [154]. Another HMM-based method in gesture recognition can identify two classes of gestures: deictic and symbolic [111].

FSM (Finite State Machine) is used for gesture recognition. Hong and Huang proposed a state based approach to gesture learning and recognition [93]. Each gesture can be defined to be an ordered sequence of states in spatial-temporal space by using Spatial Clustering and Temporal alignment. The features are the 2D image positions of the centers of the hands. From training data of a given gesture, the spatial information can be learned and grouped into segments that are automatically aligned temporally. An FSM recognizer is thus built by integrating this temporal information. Corresponding to each gesture is an FSM, the computational efficiency of which allows real-time performance.

A technique for gesture recognition is proposed that involves both physical and control models of gesture performance [130]. The path that the hand traverses while performing a gesture is modeled as a point-mass moving through air. The control model for each specific gesture is an experimentally determined sequence. These models are are used to augment a set of Kalman-filter based recognizer modules so that each filters the input data. The filter output that most closely matches the output of an unaugmented Kalman filter will be the recognized gesture.

0.4.4 Applications

There have been many implemented application systems in the hand-centered HCI. A real-time gesture-controlled interaction is described for visual modeling, analysis, and recognition of continuous dynamic hand gestures [166]. Twelve hinds of hand gestures can be recognized with accuracy of over %89 and a gesture-controlled panoramic map browser is designed and implemented as a prototype system for gesture-controlled interaction.

Recognition of sign languages (such as American Sign Language) is considered as another application that naturally employs human gestures as a means of communication.A device which could automatically translate hand gestures that represent a sign language into speech signals would have great impact on the deaf. The appearance-based modeling techniques are suited for sign language interpretation. A real-time HMM-based system for recognizing sentence-level continuous ASL (American Sign Language), by using a cap-mounted camera, can achieve %97 accuracy with an unrestricted grammar in experiments that use 40 word lexicon [144]. In a video-based continuous GSL (German Sign Language) recognition system in which HMMs are used with one model for each sign and feature vectors reflecting manual sign parameters serve as input for training and recognition, the accuracy can reach %91 on a lexicon of 97 GSL signs [65]. An experiment is carried out with 200 samples of 10 JSL (Japanese Sign Language) for the recognition of the sequences of the signed words. Each signed word is detected from the input gesture by detecting the the borders of the signed words from ordinary signed gestures of on or two hands and segmenting them. Then the segments representing the signed words are identified. The accuracies for word recognition and the sentence are %86 and %58 respectively [127].

Recognition of hand gestures as a pattern of HCI modality can fused into

another HCI modality such as speech input in order to incorporate naturalness in the design of human computer interfaces. Sharma et al have examined hand gestures made in a natural domain, that of a weather person narrating in front of a weather map [133]. This uncontrolled environment, in which the gestures of the weather man are embedded in a narration, provides abundant data to study the interaction between speech and gesture. An HMMs architecture is used for continuous gesture recognition framework and keyword spotting. And a statistical co-occurrence analysis of different gestures is conducted with a selected set of spoken keywords in order to explore the relation between gesture and speech. The experiments demonstrated the this co-occurrence analysis can improve the performance of continuous gesture recognition.

There are many other application issues about the hand gesture studies. For example, hand pose recognition in handling user hand anatomy, perspective effects, and the idiosyncrasies of individual gesture presentation [36], hand gestures for replacing the mouse in actions such as selecting, moving and resizing windows [102], recognition of 28 different hand signs by detecting hand shape, the location, and the movement [73], tracking multiple human hands [94], tracking positions of the centers and the fingertips of both hands [129], driving a computer graphics animation by the movement and configuration of the user's hand [153], detecting pointing gestures and estimating the direction of pointing and thus selecting the targets in a room and controlling the cursor on a wall screen [99], a prototype test bed system where the user can control a TV set and a lamp by using different types of hand postures [68], and etc.

0.5 Integration of Multiple Modalities

In a natural environment, some modalities are more closely related than others. For example, speech and lip movements are more closely tied than speech and hand gesture. This *closeness* can be interpreted as a ladder of different levels on which the actual fusion of different combinations of modalities depend. For the selected HCI modalities, the major topics are *when* and *how* to integrate multiple modalities. We try to explore these problems by following the framework by Sharma et al [134] and adding others' work.

0.5.1 When to Integrate the HCI Modalities

The problem of *when* to integrate the multiple modalities, in fact, determines the abstraction level at which the different modalities are fused: they are combined at the lower "raw" sensory data level or at the higher "decision" level? To address this problem, three levels of integration have been explained: data fusion, feature fusion, and decision fusion [133]. The lowest level of fusion is data fusion which involves integration of raw observations and takes place only when the observations are of the same type. Usually this does not happen in multimodal HCI integration because of of the different nature in HCI multimodality. But it is characterized by the highest level of information detail out of the three fusion types and thus assumes to be a high level of synchronization of the multimodal observations. Data fusion is susceptible to noise, specific nature and behavior of sensors.

The intermediate level, feature fusion, widely exists in HCI multimodal integration. Each stream of sensory data is analyzed for features, which are then fused. It is often used in the combination of the closely coupled and synchronized modalities such as speech and lip movement. The computational complexity of feature fusion is high because feature sets are large. Feature-level fusion possess less detailed information that data fusion but it is less sensitive to noise.

The most commonly found type of fusion in HCI is the decision fusion, the highest level of the three. It is based on the fusion of individual mode decisions or interpretations and its multimodal synchronization is the synchronization of decisions on a semantic level. The low data bandwidth of decision fusion makes its computation complexity not so expensive as that of feature fusion. Decision-level fusion is the most robust and resistant to individual sensor failure. But it can not easily recover from loss of information at lower levels of data analysis, thus unable to effectively exploit the correlation between the modality streams at the lower integration levels. It is difficult to find an optimal fusion level for a particular combination of modalities. The integration and synchronization of the modes in a natural environment may give hints to the HCI multimodal fusion.

0.5.2 How to Integrate the HCI Modalities

Sharma et al addressed this problem by at first discussing the plausible biological basis for integration and introducing a general model of fusion, and then exploring integrations in feature and decision levels [134]

In biological foundations, there are two facts relevant to multimodal fusion. 1). Evidence accruement: sensory evidence in brain structure seems to be accrued rather than averaged over different sensors inputs. 2). Contextual dependency: different signals from different brain structures are modulated and fused to induce contextual feedback. And the studies of perceptual sensory fusion suggest that there are several ways of dealing with sensory discordance: blind fusion (without any regard for individual discordances), fusion with sensor recalibration, fusion with sensor suppression, and no fusion (discordant sensors are not fused).

Based on the biological evidence about integration of multiple senses and the foundations of sensory data fusion theory, a general model of multimodal fusion for HCI is built. This fusion model assumes that each concept behind any human action is expressed through multiple action modalities and is perceived through multiple sensory modalities. Different abstractions of the observed actions (data, features, or decisions) are integrated for the multimodal fusion in that contextual knowledge can be used to constrain the search space of the problem.

Equipped with the general model, multimodal fusion can explored on different levels. Feature-level fusion is associated with two techniques. 1). FIFO (Feature In, Feature Out): the multimodal decision can be inferred with Kalman filters. Instead of performing fusion of a time series of feature vectors, the fusion is performed over a sequence of features belonging to different modalities. 2). FIDO (Feature In, Decision Out): probabilistic networks such as ANN(Artificial Neural Network) and HMM are used for FIDO fusion. As for decision-level fusion, it involves fusion of concepts(decisions) from the individual modes to form a unique multimodal concept. Its assumptions is that the basic features of the individual modes are not sufficiently correlated to be fused at the feature level. Details in the general model, feature fusion, and decision fusion can be seen at [134].

0.5.3 Multimodal HCI Systems and Applications

The substantial research interest toward developing multimodal HCI systems have resulted in many implementations of HCI multimodal integration, although most of them have not entered into market. Following is a list of computer-vision-based HCI projects.

- *Perceptive Spaces* is an unencumbered IVE (Interactive Virtual Environment) interface with a particular focus on efforts to include both the speech and gesture of the user such as real-time estimation of position, orientation, and shape of moving human head and hands, understanding of the cues of natural speech (pitch, energy, timing) [157].

- *Speech and Vision Integration for Display Control* set to develop novel and effective paradigms for human computer interface combining vision, speech, and natural language understanding: 3D-model algorithms for facial features(especially lips) extraction and tracking, hand detection and tracking; database of video/audio sequences of people making hand gestures and voice commands for controlling 3D display [150] [149].

- *StartleCam* is a wearable camera, computer, and sensing system in that the camera can be controlled via both conscious and preconscious events involving the wearer: when certain events of supposed interest to the wearer are detected through measuring the skin conductivity changes, a buffer of digital images (at this point corresponding to the *flashbulb* memory archive for the wearer) is downloaded and transmitted to mimic the wearer's own selective memory response [88].

- *KidsRoom* is a perceptually-based, interactive, narrative playspace for children where a fantasy play is staged on by using images, lighting, sound, and computer vision action recognition technology: the children's positions and actions are tracked and recognized automatically by cameras and computers and then used as input for the narration control system and finally coupled to the narrative, exploiting the context of the story [66].

- *VIGOUR* is a platform for simultaneously tracking of multiple people and recognition of their behaviors for high-level interpretation.

Through perceptual integration, different types of visual information are fused to combine the benefits of different techniques. Robust low-level visual cues such as skin color and motion are used to focus attention and facilitate real-time tracking. VIGOUR tracks behaviors using gestures and head pose to produce a high-level behavior representation for subsequent interpretation. The system is able to track three people and recognize their gestures simultaneously in real time [139].

- *3-D Visual Operating System* is a new operating system based on the 3-D display, head tracker, gaze tracker, and hand tracker subsystems. [1].

- *gaze-assisted translator* works this way: when a user reads an on-screen document in a certain foreign language and encounters difficulties (e.g. new words/phrases), the translator will detect such circumstance and deliver necessary assistance (e.g. the corresponding words/phrases in native language) in real-time [96].

- *eye interpretation engine* will distinguish, recognize, and adapt to various EM behaviors for different individuals, or different EMs for same person such as looking at screen with/without intentions [76].

- In one VMI (*Visually Mediated Interaction*) system where computer acts as an intelligent mediator between two remotely-communicating parties, head pose is exploited to impose contextual constraints for the recognition of gestures. This process was illustrated in a video-conferencing between three people and their friend overseas [135].

- *Affective Computing* is a new area for developing machine's ability to recognize human's affective states such as frustration, confusion, interest, stress, anger, and joy, where computer-vision methods and techniques can give help and assistance [126] [123] [122].

- *BlueEyes* aims at creating computational devices with the sort of perceptual abilities that humans take for granted. It combines *Affect Detection Emotion Mouse Magic Pointing* and *Pupil Finder* projects by using non-obtrusive sensing (video cameras) to extract the information such as the user's gaze, pose, hand gestures, and facial expressions [31].

0.6 Companies

1. **The Vision Corporation** [63]
   The company develops and deploys facial recognition technology, which allows computers to rapidly and accurately recognize faces.It has put on market a software named *FaceIt* that enables the broad range of applications: ID Solutions to information security, to banking, to e-commerce. FaceIt detects single or multiple faces (one-to-one matching, one-to-many matching). It works following the steps: 1). head-like object finding; 2). face detected: extracted from the background and subjected to a number of proprietary processing stages, transformed into *FACEPRINT* (an internal representation of about 84 bytes),facial identity determined by matching the live faceprint against a database of faceprints of known individuals.

2. **IBM's BlueEyes** [31]
   The IBM's project *Magic Pointing* takes advantage of the eye (by making use of gaze tracking) and pursues an approach of gaze-initiated cursor, and thus . Its current gaze-tracking technology reaches within a degree – about a half an inch on a typical screen.

3. **EyeTechDS: EyeTech digital Systems, Inc.** [21]

   - **Quick Glance:** a device (an alternative to PC mouse) that moves the cursor according to the user's eye movements (combined with on screen keyboard).
   - **Eye Science Gaze Tracking System:** an eye-tracker with sample rates up to 30 frames/sec for recording eye gaze data.

4. **Eyegaze:** [49]
   The Eyegaze system uses a video camera (infrared) to measure where the user is looking at on the screen (60 Hz and 1/4 inch accuracy).

5. **SMI: SensoMotoric Instruments** [56]
   One of the products of the company is **EyeLink**, an eye gaze tracking system. The system is a headband-mounted eye-tracking combination including: 1). two custom-built ultra-miniature high-speed cameras that take simultaneously 250 images per second; 2). a third camera that

tracks 4 IR markers for head motion compensation and gaze position tracking; 3). a powerful image processor (EyeLink Operator PC). In the tracking process, pupil and marker positions are calculated in real time to compute gaze position with extremely low noise and high resolution. Low-delay image processing can be achieved for interactive and gaze-controlled applications. The major specifications for this product is as follows:

- **Sampling rate:** 250 Hz using IR vidoe-based tracking technology simultaneously for both eyes and head position compensation.
- **Data transit delay:** 6-12 msec or typically 10 msec (time from physical eye movement until eye movement data samples are transferred to application software).
- **On-line computation** of true gaze position.
- **Gaze position tracking range:** $\pm$ 20° horizontally and $\pm$ 17°.
- **Gaze-position accuracy:** 0.5° - 1.0° average error.
- **Working distance:** 4 to 7 cm camera-to-eye distance, 40 to 140 cm display-to-eye working range.
- **On-line eye-movement parser** that detects and analyses saccades, fixations, and blinks in real time.

6. **Alphabio's Eyeputer:** [7] a video based eye-tracker which is characterized by its features and specification:

   - including a PC, 4 processing card, infrared camera, and software;
   - real-time measurement: 60Hz, 240Hz, or 480Hz sampling rate with synchronization of binocular measurements;
   - output according to Digital, Analogue, ASCII;
   - headset or installed in a car;
   - calibration on 7 (or above) points in less than 2 minutes;
   - Fov: $\pm$30°/horizontal, $\pm$20°/vertical, $\pm$45°/torsional;
   - precision: $\pm$0.1° – $\pm$0.3°, resolution: $\pm$0.016° – $\pm$0.1°;
   - prices: $25000 to $36000.

7. **ASL: Applied Science Laboratories:** [8]
   This company has manufactured a variety of eye tracking and pupillometry equipment, featured in binocular, limbus tracker, or video. The products' specifications vary according to the different models.

8. **ERICA Inc.: Eyegaze Response Interface Computer Aid:** [20]
   This company develops a technology that tracks and records a person's eye movement and pupil dilation to aid in the interaction between man and machine.

   - the ERICA system sits in a briefcase-sized box under the monitor;
   - the software fully integrated into Win95/98/NT;
   - using*only an eye*, user can gain complete control over computer system by allowing user to create documents, surf internet, play games.

9. **EYECAN's VisionKey:** [19]
   This is an assistive communications tool for people with sight but impaired motor abilities. VisionKey is operated by eye movement. It includes a Viewer mounted on glasses frames in front of one eye displaying a keyboard and containing an IR eye tracker. The price is $5000.

10. **IOTA's EyeTrace:** [46] This company provides a system for measuring, recording, and displaying binocular, horizontal, and vertical eye movements. The infrared diodes, solid state photo detectors, and integrated circuits mounted in goggles. The sampling rate is from 100/sec to 1000/sec, and the resolution can reach 0.01°.

11. **ISCAN:** [47]
    This company delivers an Eye and Target Instrumentation incorporation. It delivers the finest video based eye and (multiple) target tracking equipments of various kinds.

12. **SKALAR's 3D Image-based Eye Tracker:** [57]
    This product has integrated the programmable CMOS image sensors interfaced directly to digital processing circuitry. It has following features and specifications:

- sampling rate up to 200/sec for 3D and 400/sec for 2D;
- FoV: $\pm 90°$/horizontal, +40/-60°/vertical;
- measurement resolution $< 0.1°$;
- **software:** circle and ellipse fitting algorithms are included for pupil tracking with an implementation of the polar correlation algorithm. Additional algorithm based on artificial iris markers is included and used-defined image processing routines is also considered.

13. **faceLAB** of company *seeing machines* [55]: a head-pose and gaze direction tracking system which can be used in detecting fatigue and inattention, interactive multimedia, and human performance measurement. The product features in its real-time measurement and 60Hz update of 3D head position, eye-gaze direction, and blink detection. The accuracy in head position can reach to within 1mm and 2degrees (static) and 1mm (dynamic), and in eye-gaze upto 3degrees. It tracks without using markers, corneal illumination, just by using passive and non-intrusive measurement and allowing head movement over 30X12X50cm and 90degree head rotation.

0.7 Research Institutes and Groups

1. **CVonline: http://www.dai.ed.ac.uk/CVonline/** [86] presents: *The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision.*The Compendium is a collection of hypertext summaries on the central topics in computer vision. We have organized an index of about 700 topics. The Compendium is intended to be a clear summary of methods and applications of computer vision, organized into sections covering the main topics of practice and research. Each section will contain a number of topics, and each topic will be a hyperlink to a set of materials associated with that text. There are 19 folded subjects such as *Applications, Recognition Methods, and Visual Learning.*

2. **ISL: Interactive Systems Labs** [48]
Researchers at **ISL** are going to combine their two projects *Eye Gaze Tracking* and *Head Pose Tracking* to obtain general gaze information.

Eye Gaze Tracking: monitoring a user's eye gaze by estimating what a user is looking at on a screen based on information of the user's eye gaze. They have developed a **non-intrusive** (they consider method of using infrared light as intrusive) *facial feature tracker* that can detect and track a user's eyes in real-time (above 15 frames /sec) without the need of any special lighting or any marks on the user's face. They use iterative thresholding for searching and tracking eyes/pupils, and even lip corners and nostrils and then feed those processed image to (trained) neural networks to estimate a user's eye gaze on a screen. The accuracies for *person-dependency* and *multi person tacking* are: 1.3 and 1.9 degrees respectively.
Head Pose Tracking: The system for head pose tracking (focusing on estimating the pose of the head, e.g. the 3-D rotation and translation of the head) allows the user to freely move in the view of the camera, automatically finds and tracks facial features (eyes/pupils, nostrils, and lip corners), and can recover from tracking failures. The system runs at about 15+ frames /sec on a HP9000 using a frame grabber and a Canon VC-C1 camera. In the first place, the face is searched and tracked by using a statistical color model: 1). a stochastic model is used to characterize skin colors of faces for tracking a face; 2). a motion model to estimate image motion and to predict search window; 3). a camera model to predict and to compensate for camera motion. And then, the eyes, nostrils and lip corners are searched and tracked inside the facial region.

3. **Eye Detection and Tracking** *at gatech* [16]
   It is currently working on extracting higher level features: finding faces for multiple people and computing blink statistics, face recognition, head pose estimation, and measurement of attention/sleepiness.

4. **RSL at ANU: Robotic Systems Lab** [53]
   Stereo Face Tracking and Gaze Point Estimation: The *Safer Behicle design* (supported by *Volve*) is a system that tracks both head and eye movements by using two cameras and template matching feature tracking in real time (30Hz), that is , tracking the face, estimating the 3D head pose, estimating the gaze direction and detecting tracking failure and recovery. **Gesture Interface to Virtual Environment:**

to track the hand in a gesture-based interface to 3D environments.

5. **Pattern Recognition and Image Processing Lab at MSU:** [51]
   Its **Real-time Tracking of Face Features** is a non-intrusive real-time program that detects the eyes and nose of a moving user at a rate of between 10 and 30 frames /sec. Also It creates a base facility for other capabilities such as detecting gaze direction and facial gestures, creating face models. A skin color model is used along with geometric face features and various conditions (facial hair, 3D motion, clothing color, and eyeglass).

6. **UC Berkeley Computer Vision Group:** [11]
   Its *Recognition of Human Motions and Gestures* focuses on articulated full body motions: walking, running, dancing and other gestures. The earlier studies were about*talking lips and heads.*

7. **Computer Vision based HCI:** [10]
   A very good place for **Hand Gesture** studies.

8. **Human Hand Gesture Research (Miscellaneous:)** [30].

9. **Image Processing and Pattern Analysis Lab at Lehigh U** [35]
   Its *Gesture Recognition* is about the interpretation of hand and arm motion using multiple sensors.

10. **NASA Vision Group** [50]
    NASA Ames Research Center has a team of scientists and engineers who conduct research on human vision and visual technology for NASA missions. The researches are on *Dynamic Eye-point display, Eye Movements Metrics of Human Motion Perception and Search.*

11. **IFIP TC. 13 on HCI** [34]
    IFIP (International Federation for Information Processing) Technical Committee No 13 is focused on encouraging the development towards a science and technology of human-computer interaction.

12. **IFIP: International Federation for Information Processing** [40]
    A non-governmental, non-profit umbrella organization for national societies working in the field of information processing.

13. **Advanced Eye Interpretation Project at Stanford** [6] led by Dr. Greg Edwards, the founder of **Eyetools, Inc.**
    The combination of the Project and the business has provided the unique premium service of enabling companies to improve their ability to connect and communicate with their web site visitors by studying the browsing behaviors of online news readers.
    The research focuses on the eye tracking and develops tools to understand user behavior by modeling typical eye-movement, and aims at controlling computers with eyes. Their visualization and analysis software correlates what is displayed on the screen with keyboard and mouse activity and the user's eye-movements, thus inferring the user's mental status from the eye-movement patterns. The currently developed " Eye Interpretation Engine" parses eye-position data by analyzing eye-movement in search of recognizable patterns, and then infers a user's mental status and behavior. Their early work was the *on-screen keyboard and mouse controller* that enabled people to type with eye and control the mouse.

14. **The Camera Mouse** at the Boston College Campus School [60]
    A new technology using a video camera has been developed for disabled people who have very limited voluntary muscle control. The image from the camera is displayed on the screen and the camera is focused on the user's nose or chin. As the user moves his nose the mouse pointer is moved accordingly. Clicks and double clicks can be generated either using dwell time (hold the pointer over a spot for a certain amount of time and a click is generated; hold it there twice as long to generate a double click) or using any other clicking device.
    Another project called **EagleEyes** is also for the disabled. Instead of using camera, it is based on measuring a user's EOG (electro-oculographic potential) to allow the user to move the cursor on the screen by moving his or her eyes or head. Five surface electrodes are placed on the user's head, above and below eyes.

15. **Eye Movement Research:** a project led by Dr. Thomas Schnell at the University of Iowa, [131].
    Based on the theory that drivers are not consciously aware of their eye movements and eye movement data reveals the visual information

acquisition strategies applied by drivers, the project team have designed and implemented the eye movement data collection apparatus and data analysis software. Two basic modes have been used in the project. The first one, conducted in the specially designed instrument vehicle, involves recording and analysis of eye movements to learn more about a variety of operator information such as acquisition strategies, operator needs, and operator workload. The second mode utilizes the eye movement equipment to actually control (by *eye gaze and voice*) items in a human-machine interface.
The instrument vehicle is a 1996 Ford Taurus LX Sedan which is equipped with six video cameras for collecting and tracking different scene information. Among them is an IScan ETL 500 eye tracking system with a head mounted and/or panel mounted optics connected with an eye tracking computer for eye movement recording.
Their control activity is held in a Boeing 777 Flight Deck Simulator in which an eye movement equipment workstation is set up for processing gaze direction.
The message given on the website concerns only with brief introduction of the project and the equipment. There is no information about how the eye/gaze data is used in the simulator and how the data is recorded and analyzed.

16. **Interactive Media – Human Factors Department at Heinrich-Hertz-Institute** [39]

    The department designs and develops next generation terminal interfaces, systems, and applications that will allow the user an attractive and user-friendly access to multimedia data and interactive services. There are some of the projects in the department: gaze-controlled interactions with autostereoscopic multimedia displays, autosterescopic displays with head tracking, human factors and usability engineering, and user interfaces for cross-network and cross-service multimedia applications. The following are the relative topics to these projects:

    - *3-D Visual OS*: a new operating system based on object-oriented programming and visual programming by using the graphic editor and the advantages of 3D visualization modules, subdividing all

user-accessible software modules into the smallest units, and thus their corresponding graphic elements (gadgets) of these primitive units can be used to add animation, sound, and other database queries.

- *3D Display*: a 3D display is required for evaluation of the user's gaze and pose. It is based on the concept of directional multiplexing: different perspective views are visible only from a limited number of fixed viewing positions.
- *Head Tracker*: implementation of real-time video-based methods for detection and tracking of the user's head. Difficulties, such as variable orientations, sizes and partial occlusions of the face, the noise and poor camera resolution,are to be overcome in the two steps in the tracking process: face detection and eye location. Skin-color based approaches have been used in the former process while geometric characters of the eye are considered in the latter process.
- *Gaze Tracker*: the user's current point of fixation can be defined as the intersection of the gaze line (the line of sight of one eye) with the surface of the object being viewed. The gaze line is measured from the gaze direction (the unit vector of the gaze line) by using the cornea reflex method and the eye location determined by the head tracker) with edge detection.
- *Hand Tracker*: the skin color is used as the primary feature in hand tracking. Low-pass filtering and HMM (Hidden Markov Model) are also used to track the hand movement and to recognize very detailed hand gestures.

17. **Robotic Systems Lab**: *http://www.syseng.anu.edu.au/rsl/*, Department of Systems Engineering, The Australian National University. Its main research fields are co-operative robot systems, mobile robot navigation, active vision, robot learning and human-robot/computer interaction (HCI). It has also founded a company *Seeing Machines* [55], which develops products of human-machine interface including 3D head/eye-gaze location and tracking systems. Some of its projects are:

    - *Stereo Tracking for Head Pose and Gaze Point Estimation*: a system that tracks the pose of a person's head and estimates the

gaze-point in real-time.

- *Visual Interface for Human-Robot Interaction*: a visual interface that allows visual and tactile interaction between a human operator and a robot arm in real time.
- *A Gesture Interface to Virtual Environments*: its goal is to develop a gesture-based interface to a virtual workbench using vision systems for real-time hand-tracking.
- *Autonomous Vehicle Project*: a platform for mobile robotics and computer vision by automating the control functions: monitoring of driver vigilance and fatigue, automatic road following, automatic obstacle detection and avoidance, and autonomous control in an off-road unstructured environment.

18. **Vision Group at Queen Mary and Westfield College, University of London**: *http://www.dcs.qmw.ac.uk/research/vision/* Two projects are pose/gesture-based HCIs:

    - *ISCANIT: Recognizing Intention in Real-Time for Visually Mediated Interaction.* It undertakes research supporting VMI (Visually Mediated Interaction) and aims to develop generic view-based head and body behavioral models which will be then used to recognize intentions for active camera control–real-time tracking of multiple people. The project has contributed to three topics: *Gesture Recognition for VMI, Head Pose Estimation,* and *Tracking Body Parts under Discontinuous Motion using Probabilistic Inference.*
    - *VIGOUR: Vision Interaction based on Gestures and behaviOUR.* It is an integrated system platform, providing a GUI for both real-time image sequence input from recorded offline/captured online video streams and processed output output on the X11 display. The system has following features: all input comes from a single camera view, only the minimum of computation or model complexity is used, complementary qualities of different simple visual cues (such as color and motion, rather than a single modality) are exploited. And some of the integrated cue/perceptual modules are: skin color calculated using Gaussian mixture models in hue-saturation space, face detection for near frontal views using SVM

(Support Vector Machine), head pose estimation by using similarity measures and the condensation algorithm, gesture recognition based on the extracted features.

At [**?**], you can find many papers on recognition/tracking of head pose/gestures.

19. **BMVA:** The British Machine Vision Association and Society for Pattern Recognition [9]. It provides a forum for individuals and organizations involved in machine vision, image processing, and pattern recognition in the United Kingdom.

0.8 Books

1. **Handbook of Computer Vision Algorithms in Image Algebra** 2nd Ed. *by Joseph N. Wilson et al $ 99.95*
2. **Dynamic Vision: From Image to Face Recognition** *by Shanogang Gong et al.* $56
3. **Algorithms for Image Processing and Coomputer Vision** *by James R. Parker* $64.99 The book grew out of the team-work of five at Harvard Robotics Lab.
4. **High-level Vision: Object Recognition and Visual Cognition** *by Shimon Ullman $49.95*
5. **A Few Steps Towards 3D Active Vision** *by Thierry Vieville*
6. **Information Visualization: Perception for Design** *by Colin Ware*
7. **Readings in Human-Computer Interaction: Toward the Year 2000** By *Ronald M. Baecker et al* 2nd Ed. (p.900, including many topics)
8. **Introductory Techniques for 3-D Computer Vision** By *Emanuele Trucco* and *Alessandro Verri*, Prentice Hall
9. **Other books about CV/HCI can be found at** [22]

0.9 Journals

1. **ACM SIGGRAPH** ACM Special Interest Group on Graphics. [3]

2. **IEEE Transactions on Image Processing** [32]

3. **TPAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence** [33]

4. **IJCV: International Journal of Computer Vision** [45]

5. **Vision Research** [64]
   A journal devoted to the functional aspects of human, vertebrate and invertebrate vision and publishes experimental and observational studies, reviews, and theoretical papers firmly based upon the current facts of visual science.

6. **Human-Computer Interaction** [29]
   A journal of theoretical, empirical, and methodological issues of user science and of system design.

7. **TOCHI: ACM Transactions on Computer-Human Interaction** [4]

0.10 Conferences

1. **CHI: Conference on Human Factors in Computing Systems** *Sponcered by* **ACM***'s Special Interest Group on Computer-Human Interaction* (**ACM/SIGCHI**) [13].

   - **CHI 2001** [15].
   - **CHI'97** [14].

2. **FG2000:** Fourth IEEE International Conference on Automatic Face and Gesture Recognition [23].

3. **FG96:** Second IEEE International Conference on Automatic Face and Gesture Recognition [54].

4. **FG98:** Third IEEE International Conference on Automatic Face and Gesture Recognition [62]).

5. **ICCV: International Conference on Computer Vision**.

   - **ICCV'99:** [44].
   - **ICCV'98** [43].

6. **CVPRIP: International Conference on Computer Vision, Pattern Recognition & Image Processing**.

   - **CVPRIP2000** [42].
   - **CVPRIP'98** [41].

7. **CVPR: (IEEE Computer Society) Conference on Computer Vision and Pattern Recognition** [12].

8. **ICIP: (IEEE Signal Processing Society) International Conference on Image Processing**.

9. **ECCV: European Conference on Computer Vision.**

10. **Gesture Workshop 97** [25].

11. **Gesture Workshop 99** [26]).

12. **Gesture Workshop 2001** [26].

13. **9th International Conference on HCI** [2]

14. **Special Interest Group on Guidelines, Methods and Processes for Accessibility** [59].

15. **The incoming conferences on HCI and/or CV** [61].

16. **Information from Research and Conferences on Disabilities** [37].

17. **ACM SIGCAPH: Special Interest Group on Computers and the Physically Handicapped** [58].
    And there are two series of conferences associtiated with **SIGCAPH**: **CUU:** *Conference on Universal Usability* and **Assets** (*conference on Assistive Technologies*).

18. **Interact 2001: 8th IFIP TC.13 Conference on HCI** [38]

0.11 Miscellaneous

1. **ETRA: Eye Tracking Research & Application:**
   at *http://www.acm.org/pubs/contents/proceedings/series/etra/*

2. **ACM Digital Library:**
   at *http://www.acm.org/pubs/contents/proceedings/series/* where you can find almost all the proceedings of ACM such as **CHI** (Computer-Human Interface),**ETRA** (Eye Tracking Research and Application), **UIST** (User Interface Software and Technology)...

3. **The Computer Vision Homepage:**
   at *http://www.cs.cmu.edu/afs/cs/project/cil/ftp/html/vision.html*

4. **HCI index: at** *http://degraaff.org/hci/*

5. **http://www.eyetracking.net:** a very useful organization linking to various research agents and companies.

6. **AXIOM: http://axiom.iop.org** a search engine for technical articles.

7. **http://ibs.derby.ac.uk/emed/ : EMED – Eye Movement Equipment Database.**

8. **IEEE Computer Society: http://www.computer.org/**

Bibliography

[1] 3d visual os. http://atwww.hhi.de:80/~blick/3-D_Visual_OS/3-d_visual_os.html. Another site: http://www.hhi.de.

[2] 9th International Conference on HCI. http://uahci.ics.forth.gr. Another website: http://hcii2001.engr.wisc.edu/.

[3] ACM SIGGRAPH. http://siggraph.org/. Publication: http://www.siggraph.org/publications/.

[4] ACM Transactions on Computer-Human Interaction. http://www.acm.org/tochi/.

[5] Adaptive parallelization of model-based head tracking. http://www.cc.gatech.edu/cpl/projects/head-track/.

[6] Advanced Eye Interpretation Project at Stanford. http://eyetracking.stanford.edu/. Eyetools, Inc. at http://www.eyetools.com/.

[7] Alphabio's Eyeputer. http://www.electronica.fr/alphabio/.

[8] Applied Science Laboratories. http://www.a-s-l.com/. Technology and systems for eye tracking.

[9] British machine vision association and society for pattern recognition. http://www.bmva.ac.uk.

[10] Computer Vision based HCI. http://www.nada.kth.se/cvap/gvmdi/.

[11] Computer Vision Group at University of California, Berkeley. http://http.cs.berkeley.edu/projects/vision/vision_group.html.

[12] Conference on Computer Vision and Pattern Recognition, by IEEE Computer Society. http://www.computer.org/proceedings/cvpr.

[13] Conference on Human Factors in Computing System. http://www.acm.org/pubs/contents/proceedings/series/chi/.

[14] Conference on Human Factors in Computing System, 1997. http://www1.acm.org/sigs/sigchi/chi97/.

[15] Conference on Human Factors in Computing System, 2000. http://www.acm.org/sigs/sigchi/chi2001/.

[16] Eye detection and tracking. http://www.cc.gatech.edu/cpl/projects/pupil/index.html. at Georgia Tech.

[17] Eye Movements in Natural Tasks. http://www.cis.rit.edu/people/faculty/pelz.

[18] Eye tracking in advanced interface design. http://www.eecs.tufts.edu/~jacob/papers/barfield.html.

[19] Eyecan's visionkey. http://www.eyecan.ca/.

[20] Eyegaze Response Interface Computer Aid. http://www.ericainc.com/.

[21] EyeTech Digital Systems, Inc. http://www.eyetechds.com/.

[22] Finding books on CV/HCI on website amazon.com. http://www.amazon.com/exec/obidos/tg/browse/-/3895/107-7670577-3003769.

[23] Fourth IEEE International Conference on Automatic Face and Gesture Recognition. http://computer.org/Proceedings/fg/0580/0580toc.htm.

[24] Fourth ieee international conference on automatic face and gesture recognition 2000. http://www-prima.imag.fr/FG2000/program/index.html.

[25] Gesture Workshop 97. http://www.TechFak.Uni-Bielefeld.DE/GW97/Proceedings.html.

[26] Gesture Workshop 99. http://www.limsi.fr/GW99/.

[27] Head tracking at gatech. http://www.cc.gatech.edu/cpl/projects/head-track/.

[28] Head tracking using a textured polygonal model. http://www.cc.gatech.edu/cpl/projects/head-track/.

[29] Human-Computer Interaction. http://hci-journal.com/. Journal.

[30] Human Hand Gesture Research. http://www.cybernet.com/~ccohen/. Miscellaneous Studies on Hand Gestures.

[31] IBM's Blue Eyes. http://www.almaden.ibm.com/cs/blueeyes/.

[32] IEEE Transactions on Image Processing: A publication of the IEEE signal processing society. http://www.ieee.org/organizations/pubs/pub_preview/ip_toc.html.

[33] IEEE Transactions on Pattern Analysis and Machine Intelligence. http://www.computer.org/tpami/.

[34] IFIP TC.13 on HCI. http://www.ifip-hci.org/. IFIP Technical Committee on Human Computer Interaction.

[35] Image Processing and Pattern Analysis (about Gesture Recognition) Lab at Lehigh University. http://www.eecs.lehigh.edu/~ipal/.

[36] Inductive learning in hand pose recognition. http://www.computer.org/proceedings/fg/7713/77130078abs.htm.

[37] Information from Research and Conferences on Disabilities. http://www.dinf.org/research_conf.htm.

[38] Interact 2001: 8th IFIP TC.13 Conference on HCI. http://www.interact2001.org/.

[39] Interactive Media-Human Factors Department at Heinrich-Hertz-Institut. http://atwww.hhi.de. Another site: http://www.hhi.de.

[40] Interantioal Federation for Information Processing. http://www.ifip.or.at/.

[41] International Conference on Computer Vision, Pattern Recognition and Image Processing,1998. http://www.cs.usu.edu/Conferences/CVPRIP1998/.

[42] International Conference on Computer Vision, Pattern Recognition and Image Processing,2000. http://www.cs.usu.edu/Conferences/CVPRIP2000/.

[43] International Conference on Computer Vision,1998. http://www.umiacs.umd.edu/users/lsd/iccv/index.html.

[44] International Conference on Computer Vision,1999. http://www.cs.toronto.edu/iccv99/schedule.html.

[45] International Journal of Computer Vision. http://www.wkap.nl/journalhome.htm/0920-5691.

[46] IOTA AB Eye Trace Systems. http://www.iota.se/.

[47] ISCAN Inc. http://www.iscaninc.com/. Eye and Target Tracking Instrumentation.

[48] ISL: Interactive Systems Labs. http://www.is.cs.cmu.edu/js/. Carnegie Mellon U. and U. of Karlruhe.

[49] LC Technologies, Inc. http://www.eyegaze.com/.

[50] Nasa vision group. http://vision.arc.nasa.gov/.

[51] Pattern recognition and image processing lab at michigan state university. http://www.cse.msu.edu/~bakicve1/faces/.

[52] RSL: Robotic Systems Lab. http://www.syseng.anu.edu.au/rsl/rsl_stfacetrack.html. http://www1.volvo.com/group/research_and_technology/maineditnews/1,1901,7_101,00.htn

[53] RSL: Robotic Systems Lab. http://www.syseng.anu.edu.au/rsl/. at Australian National University.

[54] Second IEEE International Conference on Automatic Face and Gesture Recognition. http://computer.org/Proceedings/focs/7594/7594toc.htm.

[55] Seeing machines. http://www.seeingmachines.com/index.htm. A company founded by RSL at ANU: http://www.syseng.anu.edu.au/rsl/.

[56] SensoMotoric Instruments. http://www.smi.de/.

[57] SKALAR Medical. http://www.skalar.nl/. SKALAR's 3D Image-based Eye Tracker.

[58] Special Interest Group on Computers and the Physically Handicapped. http://www.acm.org/sigs/sigcaph/.

[59] Special Interest Group on Guidelines, Methods and Processes for Accessibility. http://uahci.ics.forth.gr/html/sig.html.

[60] The Camera Mouse. http://www.cs.bc.edu/~gips/CM/. at Boston College Campus School.

[61] The Incoming Conferences on HCI and/or CV. http://www-human.eie.eng.osaka-u.ac.jp/~kitamura/conference.html.

[62] Third IEEE International Conference on Automatic Face and Gesture Recognition. http://www.computer.org/proceedings/fg/8344/8344toc.htm.

[63] The vision corporation. http://www.visionics.com.

[64] Vision Research. http://www.elsevier.nl/locate/issn/00426989.

[65] Britta Bauer and Hermann Hienz. Relevant features for video-based continuous sign language recognition. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[66] Aaron Bobick and et al. The kidsroom: A perceptually-based interactive and immersive story environment. http://whitechapel.media.mit.edu/people/bobick/selected-pubs.html. KidsRoom's Website: http://whitechapel.media.mit.edu/vismod/demos/kidsroom/.

[67] E. Borovikov. Human head pose estimation by facial features location. http://www.umiacs.umd.edu/users/yab/SholarlyPaper1998/paper.html.

[68] Lars Bretzner and et al. A prototype system for computer vision based human computer interaction. http://www.nada.kth.se/cvap/abstracts/cvap251.html.

[69] Lars Bretzner and Tony Lindeberg. Qualitative multi-scale feature hierarchies for object tracking. http://www.nada.kth.se/cvap/abstracts/brelin-scsp99.html.

[70] Marco La Cascia, John Isidoro, and Stan Sclaroff. Head tracking via robust registration in texture map images. http://www.cs.bu.edu/techreports/pdf/1997-020-head-tracking.pdf.

[71] Kwang Nam Choi, Philip Worthington, and Edwin R. Hancock. Facial pose using shape-from-shading. http://www.bmva.ac.uk/bmvc/1999/contents.htm.

[72] Tanzeem Choudhury and Alex Pentland. Motion field histograms for robust
modeling of facial expressions. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker. Proceedings of the International Conference on Pattern Recognition, September 2000.

[73] Y. Cui and J.J. Weng. Hand sign recognition from intensity image sequences with complex backgrounds. http://www.computer.org/proceedings/fg/7713/7713toc.htm.

[74] Douglas DeCarlo and Dimitris Metaxas. Deformable model-based fac shape and motion analysis from images using motion residual error. http://www.cs.rutgers.edu/~decarlo/facetrack.html. In Proceedings ICCV '98, pp. 113-119, 1998.

[75] A. T. Duchowski, V. Shivashankaraiah, and T. Rawls. Binocular eye tracking in virtual reality for inspection training. In *Eye Tracking Research and Application Symposium 2000.* ACM Digital Library, 2000. ETRA Proceedings.

[76] Gregory Edwards. New software makes eyetracking viable: You can control computers with your eyes. http://eyetracking.stanford.edu.

[77] Gregory Edwards. A tool for creating eye-aware applications that adapt to changes in user behavior. http://eyetracking.stanford.edu.

[78] Ego Elagin, Johannes Steffens, and Hartmut Neven. Automatic pose estimation system for human faces based on bunch graph matching technology. http://www.computer.org/proceedings/fg/8344/8344toc.htm.

[79] S. Esaki. Quick menu selection using eye blink for eye-slaved nonverbal communicator with video-based eye-gaze detection. volume 5, pages 2322–5, 1997. Conference.

[80] C. H. Morimoto et al. Frame-rate pupil detector and gaze tracker. http://www.ime.usp.br/~hitoshi/framerate/framerate.html.

[81] C.H. Morimoto et al. Keeping an eye for hci. pages 171–176, 1999. Conference.

[82] Hans Peter Graf et al. Multi-modal systems for locating heads and faces. http://www.computer.org/proceedings/fg/7713/77130088abs.htm. http://www.research.att.com/resources/trs/TRs/96/96.5/96.5.1/96.5.1.abs.html.

[83] Ikuko Shimizu et al. Head pose determination from one image using a generic model. http://www.computer.org/proceedings/fg/8344/8344toc.htm.

[84] John M. Henderson et al. Gaze control for face learning and recognition by humans and machines. http://www.cse.msu.edu/~mahadeva/papers/book-chapter.htm. Draft of chapter to appear in: T. Shipley and P. Kellman (Eds.), From Fragments to Objects: Segmentation and Grouping in Vision.

[85] Qian Chen et al. 3d head pose estimation without feature tracking. http://www.computer.org/proceedings/fg/8344/8344toc.htm.

[86] Robert B. Fisher. Cvonline: The evolving, distributed, non-proprietary, on-line compendium of computer vision. http://www.dai.ed.ac.uk/CVonline/. Division of Informatics, University of Edinburgh.

[87] Shaogang Gong, Eng-Jon Ong, and Stephen McKenna. Learning to associate faces across views in vector space of similarities to prototypes. http://www.dcs.qmw.ac.uk/research/vision/research/published/Publications.html.

[88] J. Healey and R. W. Picard. Startlecam: A cybernetic wearable camera. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker. Proc. of the Intl. Symposium on Wearable Computers, Pittsburgh, PA, 1998.

[89] Jennifer Healey, Justin Seger, and Rosalind W. Picard. Quantifying driver stress: Developing a system for collecting and processing biometric signals in natural situations. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker. Proceedings of the Rocky-Mountian Bio-Engineering Symposium April 16-18, 1999.

[90] T. Heap and D. Hogg. Towards 3d hand tracking using a deformable model. http://www.computer.org/proceedings/fg/7713/7713toc.htm.

[91] Jochen Heinzmann and Alexander Zelinsky. 3-d facial pose and gaze point estimation using a robust real-time tracking paradigm. http://citeseer.nj.nec.com/139125.html. Third IEEE International Conference on Automatic Face and Gesture Recognition.

[92] Thomas T. Hewett and et al. Curricula for human-computer interaction. http://www1.acm.org/sigs/sigchi/cdg/. ACM Special Interest Group on Computer-Human Interaction Curriculum Development Group.

[93] Pengyu Hong and et al. Gesture modeling and recognition using finite state machines. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[94] Hitoshi Hongo and et al. Focus of attention for face and hand gesture recognition using multiple cameras. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[95] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3-d head orientation from a monocular image sequence. http://www.computer.org/proceedings/fg/7713/7713toc.htm.

[96] A. Hyrskykari, P. Majaranta, A. Aaltonen, and K. Raiha. Design issues of iDict: A gaze-assisted translation aid. In *Eye Tracking Research and Application Symposium 2000*. ACM Digital Library, 2000. ETRA Proceedings.

[97] P. Isokoski. Text input methods for eye trackers using off-screen targets. In *Eye Tracking Research and Application Symposium 2000*. ACM Digital Library, 2000. ETRA Proceedings.

[98] Qiang Ji and Xiaojie Yang. Real time visual cues extraction for monitoring driver vigilance. International Workshop on Computer Vision Systems, July 7-8, 2001, Vancouver, Canada.

[99] Nebojsa Jojic and et al. Detection and estimation of pointing gestures in dense disparity maps. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[100] Michael J. Jones and Tomaso Poggio. Multidimensional morphable models. http://citeseer.nj.nec.com/192651.html.

[101] Shinjiro Kawato and Jun Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[102] R. Kjeldsen and J. Kender. Toward the use of gesture in traditional user interfaces. http://www.computer.org/proceedings/fg/7713/7713toc.htm.

[103] Norger Kruger and et al. Estimation of face position and pose with labeled graphs. http://www.bmva.ac.uk/bmvc/1996/index.html.

[104] Volker Kruger, Sven Bruns, and Gerald Sommer. Efficient head pose estimation with gabor wavelet networks. http://www.bmva.ac.uk/bmvc/2000/contents.htm.

[105] C. Lankford. Gazetracker: Software designed to facilitate eye movement analysis. In *Eye Tracking Research and Application Symposium 2000*. ACM Digital Library, 2000. ETRA Proceedings.

[106] Chris Lankford. Effective eye-gaze input into windows. http://www.acm.org/pubs/citations/proceedings/graph/355017/p23-lankford/.

[107] Ivan Laptev and Tony Lindeberg. Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features. http://www.nada.kth.se/cvap/abstracts/cvap245.html.

[108] Mun Wai Lee and Surendra Ranganath. Pose invariant face recognition by face synthesis. http://www.bmva.ac.uk/bmvc/2000/contents.htm.

[109] Ying li Tian, Takeo Kanade, and Jeffrey F. Cohn. Dual-state parametric eye tracking. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[110] Jin Liu. Determination of the point of fixation in a head-fixed coordinate system. volume 1, pages 501–4, 1998. Conference.

[111] Sbastien Marcel and et al. Hand gesture recognition using input-output hidden markov models. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[112] Y. Matsumoto. Behavior recognition based on head pose and gaze direction measurement. volume 3, pages 2127–32, 2000. IEEE/RSJ (IROS) Conference 2000.

[113] Yoshio Matsumoto and Alexander Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[114] Keith Miller. User-centered interface design. http://www.uis.edu/ miller/se/tutorial.html.

[115] C.H. Morimoto and D. Koons. Pupil detection and tracking using multiple light sources. *Image and Vision Computing*, 18(4):331–335, 2000.

[116] Kazuhito Murakami and el al. Dynamic facial caricaturing system based on the gaze direction of gallery. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[117] Brad A. Myers. A brief history of human computer interaction technology. http://www.victoriapoint.com/hci_history.htm. ACM interactions. Vol. 5, no. 2, March, 1998. pp. 44-54.

[118] Rhys Newman and et al. Real-time stereo tracking for head pose and gaze estimation. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[119] Jeffrey Ng and Shaogang Gong. Multi-view face detection and pose estimation using a composite support vector machine across the view sphere. http://www.dcs.qmw.ac.uk/research/vision/research/published/Publications.html.

[120] Ravikanth Pappu and Paul Beardsley. A qualitative approach to classifying gaze direction. http://pappu.www.media.mit.edu/people/pappu/newpubs.html. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, April 1998.

[121] Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. http://citeseer.nj.nec.com/pavlovic97visual.html.

[122] Rosalind W. Picard. Affective medicine: Technology with emotional intelligence. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker.

[123] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker.

[124] Claudio S. Pinhanez and Aaron F. Bobick. It/i: A theater play featuring an autonomous computer graphics character. http://whitechapel.media.mit.edu/people/bobick/selectedpubs.html. Proceedings of CHI '98, Los Angeles, CA. pp. 333-334, April 1998.

[125] Thomas D. Rikert and Michael J. Jones. Gaze estimation using morphable models. http://citeseer.nj.nec.com/rikert98gaze.html. Third IEEE International Conference on Automatic Face and Gesture Recognition.

[126] Jocelyn Riseberg and et al. Frustrating the user on purpose: Using biosignals in a pilot study to detect the user's emotional state. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker. CHI98.

[127] Hirohiko Sagawa and Masaru Takeuchi. A method for recognizing a sequence of sign language words represented in a japanese sign language sentence. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[128] D. D. Salvucci. An interactive model-based environment for eye-movement protocol analysis and visulization. In *Eye Tracking Research and Application Symposium 2000*. ACM Digital Library, 2000. ETRA Proceedings.

[129] Yoichi Sato and et al. Fast tracking of hands and fingertips in infrared images for augmented desk interface. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[130] Greg S. Schmidt and Donald H. House. Towards model-based gesture recognition. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[131] Thomas Schnell. Eye movement research and eye movement apparatus at the university of iowa. http://warrior.win.ecn.uiowa.edu. Research Projects: Eye Movement Research.

[132] Andrew Sears. Hci education: Where is it headed? http://www.victoriapoint.com/hcied.htm.

[133] Rajeev Sharma and et al. Exploiting speech/gesture co-occurrence for improving continuous gesture recognition in weather narration. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[134] Rajeev Sharma, Vladimir I. Pavlovic, and Thomas S. Huang. Toward multimodal human-computer interface. http://www.ifp.uiuc.edu/IDFL/papers/paperssharma.html#sharmaab18. ieeexplore.ieee.org/iel3/5/14574/00664275.pdf.

[135] Jamie
Sherrah and Shaogang Gong. Exploiting context in gesture recognition.
http://www.dcs.qmw.ac.uk/research/vision/research/published/Publications.html.

[136] Jamie Sherrah and Shaogang Gong.
Fusion of perceptual cues for robust tracking of head pose and position.
http://www.dcs.qmw.ac.uk/research/vision/research/published/Publications.html.

[137] Jamie Sherrah and Shaogang Gong. Gesture recognition for visually mediated interaction. http://www.dcs.qmw.ac.uk/~jamie/gesture.

[138] Jamie Sherrah and Shaogang Gong. Tracking discontinuous motion using bayesian inference.
http://www.dcs.qmw.ac.uk/research/vision/research/published/Publications.html.

[139] Jamie Sherrah and Shaogang Gong. Vigour: A system for tracking and recognition of multiple people and their activities.
http://www.dcs.qmw.ac.uk/research/vision/research/published/Publications.html.

[140] Jamie Sherrah, Shaogang Gong, and Eng Jon Ong. Understanding pose discrimination in similarity space.
http://www.dcs.qmw.ac.uk/research/vision/research/published/Publications.html.

[141] Jamie Sherrah, Eng Jon Ong,
and Shaogang Gong. Estimation of human head pose using similarity measures. http://www.dcs.qmw.ac.uk/~jamie/pose/.

[142] Linda E. Sibert, Robert J.K. Jacob, and James N. Templeman. Evaluation and analysis of eye gaze interaction.
http://www.eecs.tufts.edu/~jacob/papers/nrlreport.pdf.

[143] B.A. Smith, J. Ho, W. Ark, and S. Zhai. Hand eye coordination patterns in target selection. In *Eye Tracking Research and Application Symposium 2000*. ACM Digital Library, 2000. ETRA Proceedings.

[144] Thad Starner, Joshua Weave, and Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker. PAMI; Submitted 4/26/96.

[145] Rainer Stiefelhagen and Jie Yang. Gaze tracking for multimodal human-computer. http://citeseer.nj.nec.com/141301.html. ISl: University of Karlsruhe and Carnegie Mellon University.

[146] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Towards tracking interaction between people. http://citeseer.nj.nec.com/76673.html.

[147] Jacob Strom and Alex Pentland. Face tracking using 3d models. http://www-white.media.mit.edu/~jacob/facetrack/.

[148] Kay Talmi and Jin Liu. Eye and gaze tracking for visually controlled interractive
stereoscopic displays. http://atwww.hhi.de/ liu/paper/eyegaze.pdf. http://atwww.hhi.de/ blick/Papers/eye-gaze/eye-gaze.html.

[149] Thomas and et al. Speech and vision integration for display control. http://www.ifp.uiuc.edu/IDFL/research/ressharma.html.

[150] Thomas and et al. Speech and vision integration for display control. http://www.ifp.uiuc.edu/IDFL/results/resultshuang.html.

[151] Jian-Gang Wang and Eric Sung. Gaze determination via images of irises. http://www.bmva.ac.uk/bmvc/2000/contents.htm.

[152] M. Weber and et al. Viewpoint-invariant learning and detection human heads. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[153] Andrew Wilson. Luxomatic: Computer vision for puppeteering. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker.

[154] Andrew D. Wilson and Aaron F. Bobick. Nonlinear parametric hidden markov
models. http://whitechapel.media.mit.edu/people/bobick/selected-pubs.html. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition , Santa Barbara, CA, June 1998.

[155] Andrew D. Wilson and Aaron F. Bobick. Parametric hidden markov models for gesture recognition. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker. Trans. Pattern Analysis and Machine Intelligence.

[156] Andrew D. Wilson and Aaron F. Bobick. Recognition and interpretation of parametric gesture. http://whitechapel.media.mit.edu/people/bobick/selected-pubs.html. Proceedings of International Conference on Computer Vision, Bombat, India, 1998.

[157] Christopher R. Wren and et al. Combining audio and video in perceptive spaces. http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker. 1st International Workshop on Managing Interactions in Smart Environments, December 13-14 1999, Dublin, Ireland.

[158] Ying Wu and Thomas S. Huang. Hand modeling, analysis, and recognition. http://www.ifp.uiuc.edu/~yingwu/papers/IEEE_MSP.pdf.

[159] Ying Wu and Thomas S. Huang. View-independent recognition of hand postures. http://citeseer.nj.nec.com/400733.html. In Proc. of IEEE Conf. on CVPR'2000, Vol.II, pp.88-94, Hilton Head Island, SC, 2000.

[160] Ying Wu and Thomas S. Huang. Vision-based gesture recognition: A review. http://citeseer.nj.nec.com/282466.html. Gesture Workshop.

[161] Ying Wu and Kentaro Toyama. Wide-range, person- and illumination-insensitive head orientation estimation. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[162] L-Q Xu, D. Machin, and P. Sheppard. A novel approach to real-time non-intrusive gaze finding. http://www.bmva.ac.uk/bmvc/1998/title/index.htm.

[163] Kiyotake Yachi and et al. Human head tracking using adaptive appearance models with a fixed-viewpoint pan-tilt-zoom camera. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[164] M-H Yang and N. Ahuja. Extracting gestural motion trajectories. http://citeseer.nj.nec.com/yang98extracting.html. In Proceedings of the Third International Conference on Automatic Face and Gesture.

[165] Xiaojin Zhu, Jie Yang, and Alex Waibel. Segmenting hands of arbitrary color. http://www.computer.org/proceedings/fg/0580/0580toc.htm.

[166] Yuanxin Zhu and et al. Toward real-time human-computer interaction with continuous dynamic hand gestures. http://www.computer.org/proceedings/fg/0580/0580toc.htm.