Vision Based Hand Gesture Recognition

Pragati Garg, Naveen Aggarwal and Sanjeev Sofat

***Abstract*—**With the development of ubiquitous computing, current user interaction approaches with keyboard, mouse and pen are not sufficient. Due to the limitation of these devices the useable command set is also limited. Direct use of hands as an input device is an attractive method for providing natural Human Computer Interaction which has evolved from text-based interfaces through 2D graphical-based interfaces, multimedia-supported interfaces, to fully fledged multi-participant Virtual Environment (VE) systems. Imagine the human-computer interaction of the future: A 3D-application where you can move and rotate objects simply by moving and rotating your hand - all without touching any input device. In this paper a review of vision based hand gesture recognition is presented. The existing approaches are categorized into 3D model based approaches and appearance based approaches, highlighting their advantages and shortcomings and identifying the open issues.

***Keywords*—**Computer Vision, Hand Gesture, Hand Posture, Human Computer Interface.

I. Introduction

IN the future of Steven Spielberg's *Minority Report*, Tom Cruise turns on a wall-sized digital display simply by raising his hands, which are covered with black, wireless gloves. Like an orchestra's conductor, he gestures in empty space to pause, play, magnify and pull apart videos with sweeping hand motions and turns of his wrist. Minority Report takes place in the year 2054. The touchless technology it demonstrates may arrive many decades sooner as is evident from the attention that Vision Based Interfaces have gained in the recent years.

Gestures are a powerful means of communication among humans. In fact, gesturing is so deeply rooted in our communication that people often continue gesturing when speaking on the telephone. Hand gestures provide a separate complementary modality to speech for expressing ones ideas. Information associated with hand gestures in a conversation is degree, discourse structure, spatial and temporal structure. So, a natural interaction between humans and computing devices can be achieved by using hand gestures for communication between them.

The key problem in gesture interaction is how to make hand gestures understood by computers. The approaches present can be mainly divided into "Data-Glove based" and "Vision Based" approaches. The Data-Glove based methods use sensor devices for digitizing hand and finger motions into multi-parametric data. The extra sensors make it easy to collect hand configuration and movement. However, the devices are quite expensive and bring much cumbersome experience to the users [1]. In contrast, the Vision Based methods require only a camera [2], thus realizing a natural interaction between humans and computers without the use of any extra devices. These systems tend to complement biological vision by describing artificial vision systems that are implemented in software and/or hardware. This poses a challenging problem as these systems need to be background invariant, lighting insensitive, person and camera independent to achieve real time performance. Moreover, such systems must be optimized to meet the requirements, including accuracy and robustness.

The purpose of this paper is present a review of Vision based Hand Gesture Recognition techniques for human-computer interaction, consolidating the various available approaches, pointing out their general advantages and disadvantages. Although other reviews have been written on the subsets of hand posture and gesture recognition [3], [4], [5], this one specifically relates to the vision based technique and is up-to-date. It is intended to point out the various open research issues as well as act as a starting point for anyone interested in using hand gesture recognition in their interfaces.

A. Overview

The organization of the rest of this paper is as follows. Section II highlights the various aspects of hand posture and gesture recognition technology. Section III discusses the available algorithms for hand posture and gesture recognition, discussing their advantages and shortcomings. Section IV describes various application areas of hand gesture recognition. Section V points out the open issues and Section VI concludes the paper.

II. Hand Posture and Gesture Recognition Technology

The human hand has a complex anatomical structure consisting of many connected parts and joints, involving complex relations between them providing a total of roughly 27 degrees of freedom (DOFs) [6]. User Interface development requires a sound understanding of human hand's anatomical structure in order to determine what kind of postures and gestures are comfortable to make. Although hand postures and gestures are often considered identical, the distinctions between them need to be cleared. Hand posture is a static hand pose without involvement of movements. For

Pragati Garg is with Punjab Engineering College, CSE Deptt., Sector 12, Chandigarh, 160012, India (e-mail: pragati.garg@gmail.com).

Naveen Aggarwal is with University Institute of Engineering and Technology, CSE Deptt, Panjab University, Sector 25, Chandigarh,160025, India. (e-mail navagg@gmail.com).

Sanjeev Sofat is with Punjab Engineering College, CSE Deptt. Sector 12, Chandigarh 160012, India (e-mail: sanjeevsofat@pec.ac.in).

example, making a fist and holding it in a certain position is a hand posture. Whereas, a hand gesture is defined as a dynamic movement referring to a sequence of hand postures connected by continuous motions over a short time span, such as waving good-bye. With this composite property of hand gestures, the problem of gesture recognition can be decoupled into two levels- the low level hand posture detection and the high level hand gesture recognition.

In vision based hand gesture recognition system, the movement of the hand is recorded by video camera(s). This input video is decomposed into a set of features taking individual frames into account. Some form of filtering may also be performed on the frames to remove the unnecessary data, and highlight necessary components. For example, the hands are isolated from other body parts as well as other background objects. The isolated hands are recognized for different postures. Since, gestures are nothing but a sequence of hand postures connected by continuous motions, a recognizer can be trained against a possible grammar. With this, hand gestures can be specified as building up out of a group of hand postures in various ways of composition, just as phrases are build up by words. The recognized gestures can be used to drive a variety of applications (Fig.1).

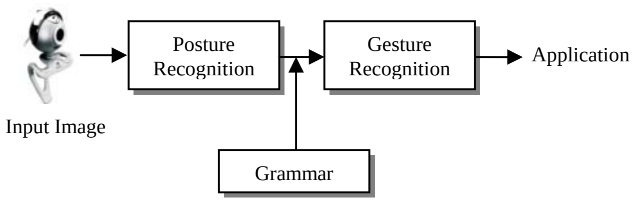

Fig. 1. Hand Gesture Recognition Process

III. Approaches for Hand Posture and Gesture Recognition

The approaches to Vision based hand posture and gesture recognition can be divided into two categories – 3 D hand model based approaches and appearance based approaches [7].

A. *3 D Hand Model based Approach*

Three dimensional hand model based approaches rely on the 3 D kinematic hand model with considerable DOF's, and try to estimate the hand parameters by comparison between the input images and the possible 2 D appearance projected by the 3-D hand model. Such an approach is ideal for realistic interactions in virtual environments.

One of the earliest model based approaches to the problem of bare hand tracking was proposed by Rehg and Kanade [8]. This article describes a model-based hand tracking system, called DigitEyes, which can recover the state of a 27 DOF hand model from ordinary gray scale images at speeds of up to 10 Hz. The hand tracking problem is posed as an inverse problem: given an image frame (e.g. edge map) find the underlying parameters of the model. The inverse mapping is non-linear due to the trigonometric functions modeling the joint movements and the perspective image projection. A key observation is that the resulting image changes smoothly as the parameters are changed. Therefore, this problem is a promising candidate for assuming local linearity. Several iterative methods that assume local linearity exist for solving non-linear equations (e.g. Newton's method). Upon finding the solution for a frame the parameters are used as the initial parameters for the next frame and the fitting procedure is repeated. The approach can be thought of as a series of hypotheses and tests, where a hypothesis of model parameters at each step is generated in the direction of the parameter space (from the previous hypothesis) achieving the greatest decrease in mis-correspondence. These model parameters are then tested against the image. This approach has several disadvantages that has kept it from real-world use. First, at each frame the initial parameters have to be close to the solution, otherwise the approach is liable to find a suboptimal solution (i.e. local minima). Secondly, the fitting process is also sensitive to noise (e.g. lens aberrations, sensor noise) in the imaging process. Finally, the approach cannot handle the inevitable self-occlusion of the hand.

In [9] a deformable 3D hand model is used (Fig 2). The model is defined by a surface mesh which is constructed via PCA from training examples. Real-time tracking is achieved by finding the closest possibly deformed model matching the image. The method however, is not able to handle the occlusion problem and is not scale and rotation invariant.

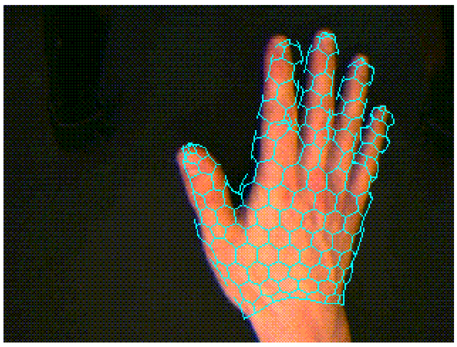

Fig. 2. Snapshot of 3 D Tracker in action

In [10] a model-based method for capturing articulated hand motion is presented. The constraints on the joint configurations are learned from natural hand motions, using a data glove as input device. A sequential Monte Carlo tracking algorithm, based on importance sampling, produces good results, but is view-dependent, and does not handle global hand motion.

Stenger et al. [11] presented a practical technique for model based 3D hand tracking (Fig 3). An anatomically accurate hand model is built from truncated quadrics. This allows for the generation of 2D profiles of the model using elegant tools from projective geometry, and for an efficient method to handle self-occlusion. The pose of the hand model is estimated with an Unscented Kalman filter (UKF) [12], which minimizes the geometric error between the profiles and edges extracted from the images. The use of the UKF permits higher frame rates than more sophisticated estimation methods such as particle filtering, whilst providing higher accuracy than the extended Kalman filter. For a single camera the tracking algorithm operates at a rate of 3 frames per second on a Celeron 433MHz machine. However, the computational complexity grows linearly with the number of cameras, making the system non-operatable in real time environments.

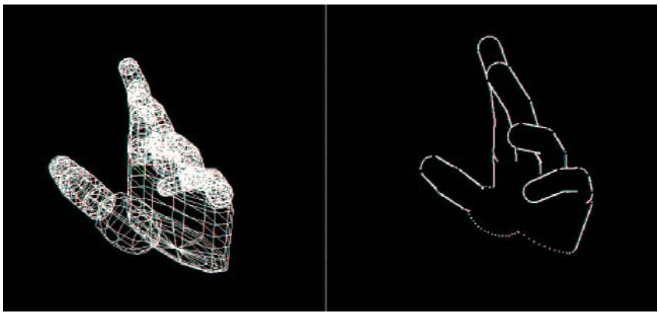

Fig. 3. The 3 D model (left) and its generated contour (right)

More recent efforts have reformulated the problem within a Bayesian (probabilistic) framework [13]. Bayesian approaches allow for the pooling of multiple sources of information (e.g. system dynamics, prior observations) to arrive at both an optimal estimate of the parameters and a probability distribution of the parameter space to guide future search for parameters. On contrary to Kalman filter approach, Bayesian approaches allow non-linear system formulations and non-Gaussian (multi-modal) uncertainty (e.g. caused by occlusions) at the expense of a closed-form solution of the uncertainty. A potential problem with the approach is that certain independent assumptions of the underlying probabilistic distribution are made, for computational tractability reasons that may not hold in practice. Also, it is a computationally expensive approach.

Three dimensional hand model based approaches offer a rich description that potentially allows a wide class of hand gestures. However, as the 3D hand models are articulated deformable objects with many DOF's, a very large image database is required to cover all the characteristic shapes under different views. Another common problem with model based approaches is the problem of feature extraction and lack of capability to deal with singularities that arise from ambiguous views.

B. *Appearance based approaches*

Appearance based approaches use image features to model the visual appearance of the hand and compare these parameters with the extracted image features from the video input. Generally speaking, appearance based approaches have the advantage of real time performance due to the easier 2 D image features that are employed.

There have been a number of research efforts on appearance based methods in recent years. A straightforward and simple approach that is often utilized is to look for skin colored regions in the image [14], [15]. Although very popular, this has some drawbacks. First, skin color detection is very sensitive to lighting conditions. While practicable and efficient methods exist for skin color detection under controlled (and known) illumination, the problem of learning a flexible skin model and adapting it over time is challenging. Secondly, obviously, this only works if we assume that no other skin like objects are present in the scene. Lars and Lindberg [16] used scale-space color features to recognize hand gestures. Their gesture recognition method is based on feature detection and user independence, but the authors showed real time application only with no other skin coloured objects present in the scene. So, although, skin color detection is a feasible and fast approach given strictly controlled working environments, it is difficult to employ it robustly on realistic scenes.

Another approach is to use the eigenspace for providing an efficient representation of a large set of high-dimensional points using a small set of basis vectors. The eigenspace approach seeks an orthogonal basis that spans a low-ordered subspace that accounts for most of the variance in a set of examplar images. To reconstruct an image in the training set a linear combination of the basis vectors (images) are taken, where the coefficients of the basis vectors are the result of projecting the image to be reconstructed on to the respective basis vectors. In [17] the authors present an approach for tracking hands by an eigenspace approach. The authors provide three major improvements to the original eigenspace approach formulation, namely, a large invariance to occlusions, some invariance to differences in background from the input images and the training images, and the ability to handle both small and large affine transformations (i.e. scale and rotation) of the input image with respect to the training images. The authors demonstrate their approach with the ability to track four hand gestures using 25 basis images. For a small set of gestures this approach may be sufficient. With a large gesture vocabulary (e.g. American Sign Language) the space of views is large, this poses a problem for collecting adequate training sets and more seriously the compactness in the subspace required for efficient processing may be lost.

Recently there has been increased interest in approaches working with local invariant features [18], [19], [20], [21]. In [18], AdaBoost learning algorithm is used with Scale invariant feature transform- a histogram representing gradient orientation and magnitude information within a small image patch. Using the sharing feature concept, efficiency of 97.8% is achieved. However, different features such as contrast context histogram need to be studied and applied to accomplish hand posture recognition in real time. In [19], [20], [21] Haar like features are used for the task of hand detection. Haar like features focus more on the information within a certain area of the image rather than each single pixel. To improve classification accuracy and achieve real-time performance, AdaBoost learning algorithm that can adaptively select the best features in each step and combine them into a strong classifier can be used. The training algorithm based on AdaBoost learning algorithm takes a set of "positive" samples, which contain the object of interest and a set of "negative" samples, i.e., images that do not contain objects of interest. During the training process, distinctive Haar-like features are selected to classify the images containing the object of interest at each stage. Originally for the task of face tracking and detection, Viola and Jones [22] proposed a statistical approach to handle the large variety of human faces. In their algorithm, the concept of “integral image” is used to compute a rich set of Haar-like features. Compared with other approaches, which must operate on multiple image scales, the integral image can achieve true scale invariance by eliminating the need to compute a multiscale image pyramid and significantly reduces the image processing time. The Viola and Jones algorithm is approximately 15 times faster than any previous approaches while achieving accuracy that is equivalent to the best published results [22]. However, training with this method is computationally expensive prohibiting the evaluation of many hand appearances for their suitability to detection.

The idea behind the invariant features is that, if it is possible to identify characteristic points or regions on objects, an object can be represented as assembly of these regions i.e. rather than modeling the object as a whole, one models it as a collection of characteristic parts. This has the advantage that partial occlusions of an object can be handled easily, as well as considerable deformations or changes in viewpoint. As long as a sufficient number of characteristic regions can be identified, the object may still be found. Therefore, these approaches seem rather promising for the task of real time hand detection.

IV. APPLICATION AREAS

Hand gesture recognition finds applications in varied domains including virtual environments, smart surveillance, sign language translation, medical systems etc. The following section gives a brief overview of few of the application areas.

Hand gestures can be used for analyzing and annotating video sequences of technical talks. Such a system is presented in [23]. Speaker’s gestures such as pointing or writing are automatically tracked and recognized to provide a rich annotation of the sequence that can be used to access a condensed version of the talk. Given the constrained domain a simple ``vocabulary" of actions is defined, which can easily be recognized based on the active contour shape and motion. The recognized actions provide a rich annotation of the sequence that can be used to access a condensed version of the talk from a web page.

For achieving natural human computer interaction for virtual environments, Pavlovic and Berry [24] integrated controlling gestures into the virtual environment BattleField. In this system hand gestures are used not only for navigating the VE, but also as an interactive device to select and move the virtual objects in the BattleField. Another system where hand gestures serve as the input and controlling device of the virtual environment is presented in [25]. This system permits interactive modeling of biopolymers by linking a 3D molecular graphics and molecular dynamics simulation program, with input being provided using hand gestures.

Also, hand gestures are an attractive method for communication with the deaf and dumb. One of the most structured sets of gestures is those belonging to any of the several sign languages. In sign language, each gesture already has assigned meaning, and strong rules of context and grammar may be applied to make recognition tractable. Thad Starner et al. [26] described an extensible system which used a single color camera to track hands in real time and interpret American Sign Language (ASL).

When robots are moved out of factories and introduced into our daily lives they have to face many challenges such as cooperating with humans in complex and uncertain environments or maintaining long-term human-robot relationships. Hand gesture recognition for robotic control is presented in [18, 27]

Recently, Hongmo Je et al. [28] proposed a vision based hand gesture recognition system for understanding musical time pattern and tempo that was presented by a human conductor. When the human conductor made a conducting gesture, the system tracked the COG of the hand region and encoded the motion information into the direction code.

Another important application area is that of vehicle interfaces. A number of hand gesture recognition techniques for human vehicle interface have been proposed time to time [29,30]. The primary motivation of research into the use of hand gestures for in-vehicle secondary controls is broadly based on the premise that taking the eyes off the road to operate conventional secondary controls can be reduced by using hand gestures.

The healthcare area has also not been left untouched by this technological wave. Wachs et al. [31] developed a gesture based tool for sterile browsing of radiology images. A sterile human-machine interface is of supreme importance because it is the means by which the surgeon controls medical information, avoiding patient contamination, the operating room and the other surgeons. The gesture based system could replace touch screens now used in many hospital operating rooms which must be sealed to prevent accumulation or spreading of contaminants and requires smooth surfaces that must be thoroughly cleaned after each procedure - but sometimes aren't. With infection rates at hospitals now at unacceptably high rates, the hand gesture recognition system offers a possible alternative.

In June 2008, Toshiba formally announced the details of Qosmio laptops offering what may be the first integration of gesture recognition and day-to-day computing [32]. With Toshiba's media center software, users can pause or play videos and music by holding an open palm up to the screen. Make a fist, and your hand functions as a mouse, pulling a cursor around the screen. Flip your thumb up and down to click

In summary, hand gesture recognition could be used in many settings in the future. The decreasing hardware and processing costs is a significant factor making gesture recognition more practical for widespread use.

V. OPEN ISSUES

Vision based hand gesture recognition is still an important area of research because the available algorithms are relatively primitive when comparing with mammalian vision. A main problem hampering most approaches is that they rely on several underlying assumptions that may be suitable in a controlled lab setting but do not generalize to arbitrary settings. Several common assumptions include: assuming high contrast stationary backgrounds and ambient lighting conditions. Also, recognition results presented in the literature are based on each author's own collection of data, making comparisons of approaches impossible and also raising suspicion on the general applicability. Moreover, most of the methods have a limited feature set. The latest trend for hand gesture recognition is the use of AI to train classifiers, but the training process usually requires a large amount of data and choosing features that characterize the object being detected is a time consuming task.

Another problem that remains open is recognizing the temporal start and end points of meaningful gestures from continuous motion of the hand(s). This problem is sometimes referred to as "gesture spotting" or temporal gesture segmentation. The ways to reduce the training time and develop a cost effective real time gesture recognition system which is robust to environmental and lighting conditions and does not require any extra hardware poses grand yet exciting research challenge.

VI. CONCLUSION

In today's digitized world, processing speeds have increased dramatically, with computers being advanced to the levels where they can assist humans in complex tasks. Yet, input technologies seem to cause a major bottleneck in performing some of the tasks, under-utilizing the available resources and restricting the expressiveness of application use. Hand Gesture recognition comes to rescue here. Computer Vision methods for hand gesture interfaces must surpass current performance in terms of robustness and speed to achieve interactivity and usability. A review of vision-based hand gesture recognition methods has been presented. Considering the relative infancy of research related to vision-based gesture recognition remarkable progress has been made. To continue this momentum it is clear that further research in the areas of feature extraction, classification methods and gesture representation are required to realize the ultimate goal of humans interfacing with machines on their own natural terms.

REFERENCES

[1] A. Mulder, "Hand gestures for HCI", *Technical Report 96-1*, vol. Simon Fraster University, 1996

[2] F. Quek, "Towards a Vision Based Hand Gesture Interface", pp. 17-31, in *Proceedings of Virtual Reality Software and Technology,* Singapore, 1994.

[3] Ying Wu, Thomas S Huang, " Vision based Gesture Recognition : A Review", Lecture Notes In Computer Science; Vol. 1739 , *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, 1999

[4] K G Derpains, "A review of Vision-based Hand Gestures", Internal Report, Department of Computer Science. York University, February 2004

[5] Richard Watson, "A Survey of Gesture Recognition Techniques", Technical Report TCD-CS-93-11, Department of Computer Science, Trinity College Dublin, 1993.

[6] Y. Wu and T.S. Huang, "Hand modeling analysis and recognition for vision-based human computer interaction", *IEEE Signal Processing Mag. – Special issue on Immersive Interactive Technology*, vol.18, no.3, pp. 51-60, May 2001

[7] H. Zhou, T.S. Huang, "Tracking articulated hand motion with Eigen dynamics analysis", In Proc. Of *International Conference on Computer Vision*, Vol 2, pp. 1102-1109, 2003 .

[8] JM Rehg, T Kanade, "Visual tracking of high DOF articulated structures: An application to human hand tracking" ,In Proc. *European Conference on Computer Vision*, 1994

[9] A. J. Heap and D. C. Hogg, "Towards 3-D hand tracking using a deformable model", In *2nd International Face and Gesture Recognition Conference*, pages 140–145, Killington, USA,Oct. 1996.

[10] Y. Wu, L. J. Y., and T. S. Huang. "Capturing natural hand Articulation". In Proc. 8th *Int. Conf. on Computer Vision*, volume II, pages 426–432, Vancouver, Canada, July 2001.

[11] B. Stenger P. R. S. Mendonc¸a R. Cipolla, "Model-Based 3D Tracking of an Articulated Hand" , In proc. *British Machine Vision Conference*, volume I, Pages 63-72, Manchester, UK, September 2001

[12] [Online] Available from: <http://en.wikipedia.org/wiki/Kalman_filter> [11.09.2008]

[13] B. Stenger, A. Thayananthan, P.H.S. Torr, R. Cipolla , "Model-based hand tracking using a hierarchical Bayesian filter" , *IEEE transactions on pattern analysis and machine intelligence*, September 2006

[14] Elena Sánchez-Nielsen, Luis Antón-Canalís, Mario Hernández-Tejera, "Hand Getsure recognition for Human Machine Intercation", In Proc. *12th International Conference on Computer Graphics, Visualization and Computer Vision : WSCG,* 2004

[15] B. Stenger, "Template based Hand Pose recognition using multiple cues", In Proc. *7th Asian Conference on Computer Vision : ACCV* 2006

[16] Lars Bretzner, Ivan Laptev and Tony Lindeberg, "Hand gesture recognition using multiscale color features, hieracrchichal models and particle filtering", in Proceedings of *Int. Conf. on Automatic face and Gesture recognition*, Washington D.C., May 2002

[17] M. Black, & Jepson, " Eigen tracking: Robust matching and tracking of articulated objects using a view-based representation", In *European Conference on Computer Vision*, 1996.

[18] C C Wang, K C Wang, "Hand Posture recognition using Adaboost with SIFT for human robot interaction", Springer Berlin, ISSN 0170-8643, Volume 370/2008

[19] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in Proc. *IEEE Int. Conf. Image Process*., 2002, vol. 1, pp. 900–903.

[20] Andre L. C. Barczak, Farhad Dadgostar, "Real-time hand tracking using a set of co-operative classifiers based on Haar-like features", *Res. Lett. Inf. Math. Sci*., 2005, Vol. **7**, pp 29-42

[21] Qing Chen , N.D. Georganas, E.M Petriu, "Real-time Vision based Hand Gesture Recognition Using Haar-like features", *IEEE Transactions on Instrumentation and Measurement* -2007

[22] P. Viola and M. Jones, "Robust real-time object detection," Cambridge Res. Lab., Cambridge, MA, pp. 1–24, Tech. Rep. CRL2001/01, 2001.

[23] Ju, S., Black, M., Minneman, S., Kimber, D, "Analysis of Gesture and Action in Technical Talks for Video Indexing", *IEEE Conf. on Computer Vision and Pattern Recognition*, *CVPR 97* ,1997

[24] Berry,G.: Small-wall, "A Multimodal Human Computer Intelligent Interaction Test Bed with Applications", Dept. of ECE, University of Illinois at Urbana-Champaign, MS thesis (1998)

[25] Zeller, M., et al, " A Visual Computing Environment for Very Large Scale Biomolecular Modeling", Proc. *IEEE Int. Conf. on Application-specific Systems, Architectures and Processors* (ASAP), Zurich, pp. 3-12, 1997

[26] Thad Starner and Alex Pentland, "Real time American Sign Language Recognition from Video using Hidden Markov Models", Technical Report 375, MIT Media Lab,1995

[27] A Malima, E Ozgur, M Cetin, "A fast algorithm for Vision based hand gesture recognition for robot control", *14th IEEE conference on Signal Processing and Communications Applications*, April 2006

[28] Hongmo Je Jiman Kim Daijin , "Hand Gesture Recognition to Understand Musical Conducting Action" *16th IEEE International Symposium on Robot and Human interactive Communication, 2007, RO-MAN*, 2007, pg 163-168

[29] Dong Guo Yonghua, " Vision-Based Hand Gesture Recognition for Human-Vehicle Interaction", *International Conference on Control, Automation and Computer Vision*, 1998

[30] Pickering, Carl A. Burnham, Keith J. Richardson, Michael J. Jaguar , "A research Study of Hand Gesture Recognition Technologies and Applications for Human Vehicle Interaction", *3rd Conference on Automotive Electronics*, 2007

[31] Juan P. Wachs , Helman I. Stern, Yael Edan, Michael Gillam, Jon Handler, Craig Feied, Mark Smith, "A Gesture-based Tool for Sterile Browsing of Radiology Images", Journal of the American Medical Informatics Association (2008; 15:321-323, DOI 10.1197/jamia.M24)

[32] [Online] Available from: <http://www.xbitlabs.com/news/mobile/display/20080624203611_Toshiba_Formally_Unveils_Notebooks_with_SpursEngine_Chip.html> [11.09.2008]