Project ref. no.	*IST*-2000-26434
Project acronym	**FGNET**
Project full title	*Face and Gesture Recognition Working Group*
Security (distribution level)	*Public*
Contractual date of delivery	*15-01-2004*
Actual date of delivery	*27.01.2004*
Deliverable number	*D2.3*
Deliverable name	Workshop 3 Report
Type	*Report*
Status & version	*Version 1.0*
Number of pages	*25*
WP contributing to the deliverable	*WP2*
WP / Task responsible	*WP2 Foresight Workshop / TUM*
Other contributors	-
Author(s)	Frank Wallhoff
EC Project Officer	*Phillipe Gelin*
Keywords	*Face Recognition, Gesture Recognition, Human Machine Interaction, Foresight Report*
Abstract (for dissemination)	The third FGNet foresight workshop was held at the Elias Beach Hotel in Limassol, Cyprus from the 28. - 29. August 2003. The topic was "Human Machine Interaction". This document gives an overview of the workshop's content and a series of scenarios how people will interact with machines in the future.

FGNet - 3rd Foresight Report

Date of preparation: 27 Jan 2004

Content List:

1 Introduction

One of the major objectives of the FGNet Network of Excellence in Face & Gesture Recognition is the organization of foresight workshops, where the FGNet members and invited experts get together in order to define visions of possible future scenarios enabled by intelligent methods in face and gesture recognition.

The third and last of these FGNet foresight workshop series was arranged by Frank Wallhoff and Gerhard Rigoll, Munich University of Technology. It was hosted by Andreas Lanitis from the Cyprus College at the Elias Beach Hotel in Limassol from the 28. - 29. August 2003. This document reports about the content and outcome of this workshop. The outline of this report is as follows: First, a brief introduction into the workshop topic “Human Machine Interaction” is given. The subsequent sections contain summaries of the talks presented by the invited speakers and related projects of the network members, which were mainly presented at the first day. The following sections describe the foresight visions and roadmaps that were defined by two different working groups during the last afternoon of the workshop. In the final section the major conclusions are summarized. The appendix contains the final programme of the workshop and a list of the participants.

2 Gerhard Rigoll: "Introduction into the workshop topic “Human Machine Interaction”

It was decided by the organizers that each face & gesture recognition workshop should be dedicated to a special topic. Talks and discussions at the workshop should be mainly centered around, but not strictly limited to this workshop topic.

The topic of the last workshop was selected to be “Human Machine Interaction”. In contrast to last year´s topic, which was "Interacting people", this topic addresses a broad field of possible scenarios. These scenarios were introduced as:

- information kiosks
- command of electronic devices
- smart rooms
- robotics
- communication in noisy environments
- communicaqtion in mobile enviroments (pedestrian, automotive)
- virtual and mixed reality environments
- sign language
- games

In conjunction with this topic there will be a wide field of involved disciplines, such as:

- tracking
- gesture recognition (body, arm, hand, static & dynamic)

- action recognition, such as pointing
- emotion recognition
- person identification for personalization of human-machine interfaces
- speech recognition
- fusion with other modalities (acoustic, haptic, ….)

To solve the technical requirements the following involved algorithms can be considered:

- segmentation
- motion
- Kalman filters
- template matching
- elastic models
- HMMs
- Neural nets, SVMs, …

The topic "Human Machine Interaction" is also known under slightly different synonyms, such as "Man Machine Interaction". It has been considered by most workshop participants that "Human Machine Interaction" is not a definite standard yet, but seems to be the most comprehensive one. This is due to the fact, that it contains the interaction between human beings, without gender issues and any kind of machine, including robots, computers and so forth.

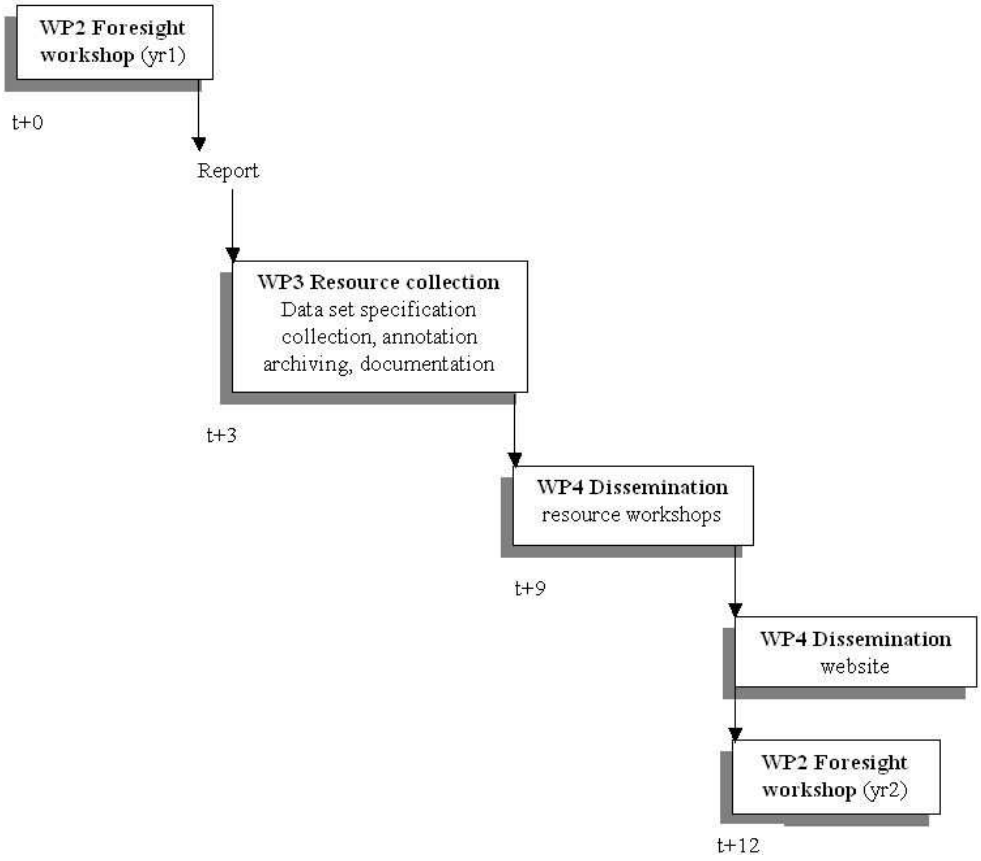
Already the discussion about the meaning of the workshop's topic shows, that this field involves several face and gesture activities. This implies that interaction is not only limited to computers, by using keyboards or other haptics, but may also be understood in all forms of interaction with machines.

3 Tim Cootes: "Message from the project-coordinator, Overview over FGNet"

During this topic on the agenda, the presenter and project coordinator Tim Cootes briefly summarized the aims and goals of the project. He also shortly introduced the involved project partners and the project plan until now.

The two previous foresight workshops and their outcome were also shortly recapitulated:

- 1st Foresight Workshop
  - INRIA, Grenoble, 1-2nd November 2001
  - Topic: “Smart Spaces”–homes,meeting rooms,offices,streets etc
  - Considerable F&G interest: surveillance, pointing/commands, people detection/tracking, face detection/recognition etc
  - Collected databases: Hand posture database, Moving People
  - Dissemination workshops: PETS and PETS 2002 at ECCV



- 2nd Foresight
  - Topic: “Interacting People”
  - IDIAP, Martigny,12-13 September 02
  - Discussion of such scenarios as:
    - Face to face meetings
    - Video conferences
    - Seminars and lectures
    - Negotiations
    - Sign language conversations
  - Collected databases: Smart Meeting Room Dataset
  - Dissemination workshop: PETS-ICVS 2003 in Graz

The presentation closed with the conclusion to find ideas for possible datasets, that can be used as a standard at dissemination workshops.

4 Summary of the invited talks

After the organizational opening and the introduction, five invited talks were presented, which are briefly summarized separately in the following paragraphs. The talks were all placed before the definition of the roadmaps in the agenda. They are all closely related to the theme of the workshop and mostly took place on the first day. The speakers were selected to represent multiple disciplines within the research community related to this topic. The

following sections contain summarizations over the contents of the given presentations. The figures in the following sections were provided by the corresponding author.

4.1 Ipke Wachsmuth: "Embodied Communication"

The first invited talk was presented by Prof. Ipke Wachsmuth, head of the *Artificial Intelligence Group, Faculty of Technology, University of Bielefeld,* entitled "Embodied Communication".

Abstract
Cognition arose in living organisms, in nature it is inseparable from a body, and only makes sense in a body. Likewise, natural communication and human language developed in intimate connection with body. When a person speaks, not only symbols are transmitted, but the whole body is in continuous motion. While speaking we can indicate the size and shape of an object by a few handstrokes, direct attention to a referenced object by pointing or gaze, and modify what we communicate by emotional facial expression. The meanings we transmit this way are multimodally encoded and strongly situated in the present context.

Embodied communication is the term meant to refer to such, often spontaneous, behavioral phenomena. Over and above symbolic communication they may convey meanings in a form which is not part of a conventionalized code but nevertheless understandable. An iconic gesture, like the one illustrated in the left figure, can serve to represent and communicate a mental image in an embodied form (McNeill, 1992). Such a gestural sign obtains meaning by iconicity, i.e. a pictorial similarity between itself and its imagined referent. An emotional expression communicates an emotional state which in its subtlety can hardly be conveyed by symbols but enhances the representational power of symbolizations.

Communication models that emphasize symbolic information transfer neglect the decisive role of non-symbolic qualities which are especially present in face-to-face communication. The cognitive modeling challenge is to devise theoretically grounded and empirically guided operational models that specify how mental processes and embodiment work together in communication.

Artificial Humanoid Agents
A growing body of work in AI and agent research – in areas like facial expression robots or embodied conversational agents – takes up questions that can be related to embodied communication in a technical way. With the artificial humanoid agent MAX under development at the University of Bielefeld we explore to what extent embodied communication can be realized by an artificial agent embodied in virtual reality. Clearly such an agent does not have a body in the physical sense, but it can be equipped with verbal conversational abilities, and employ its virtual body to express non-linguistic communication qualities such as gesture (Kopp & Wachsmuth, 2002). Equipped with a synthetic voice and an articulated body and face, Max is able to speak and gesture, and to mimic emotions. By means of microphones and tracker systems, Max can also "hear" and "see" and is able to process spoken instructions and gestures.

One of our current research challenges pertains to the question of how far Max can imitate iconic gestures demonstrated by a human communication partner. Iconic gestural movements are assumed to derive from imagistic representations in working memory, which are transformed into patterns of control signals executed by motor systems (e.g., de Ruiter, 2000). Could an artificial agent construct a "mental image" of shape from an observed iconic gesture and reexpress – *reembody* – it by way of iconic gestures? Another research challenge is emotion. Could an artificial agent express emotions related to internal parameters that are themselves influenced by external and internal events?

Conclusion

We have used examples to support the research importance of embodiment in communication. A fuller investigation would certainly involve many further aspects, e.g., rhythmic entrainment between communication partners, and so forth. Our research is led by the expectation that the construction and test of an "artificial communicator" will help to reach a more profound understanding of embodied communication. Finally, as embodiment plays such an important part in human communication, embodied communication should also have great impact in human interface research.

4.2 Joëlle Coutaz: "Distributed User Interfaces and Multi-surface Interaction"

Prof. Joëlle Coutaz, head of the *CLIPS-IMAG laboratory at the Joseph Fourier University* presented the second invited talk which was "Distributed User Interfaces and Multi-surface Interaction".

Abstract

Physical surfaces are pervasive and serve many purposes. Digital computation is a powerful source of functional support. However, it has been confined to the augmentation of single objects only. In this talk, we are interested in the combination of physicality with computation in the context of multiple objects. We propose the notion of multi-surface interaction as a unifying paradigm for reasoning about both emerging distributed UI's and known interaction techniques such as GUIs, tangible UIs, and manipulatable UIs. Multi-surface interaction through an ontology that feeds into the design of sound foundational software for the development of next generation user interfaces is defined.

Content

The presentation is about user interfaces (UIs) using pervasive surfaces. These surfaces contain projected information and recognize actions being performed of these mixed reality surfaces. These can be interpreted in several sizes, such as in the large - wall size surfaces - as well as in the small -miniature surfaces. In conjunction with this talk the expression "Human Computer Interaction" was emphasized. One of the key ideas is that UI is not restricted to one surface, but can be extended to several ones. User interfaces can be **distributed** across multiple surfaces. They furthermore are capable of **migrating** between surfaces.

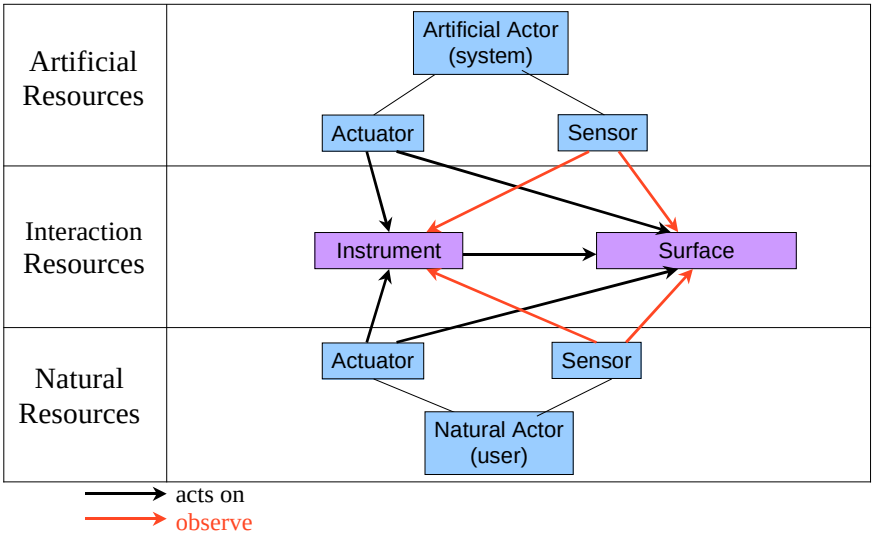
Multi surface interaction can have the following properties: wearable, graspable, movable, reachable, perennial, traversable (rain curtain), etc. Furthermore it can be composable, augmentable with computation and manipulatable by humans

Ontology for Multi-surface Interaction

The following image unifies a framework for reasoning about:

- Emerging distributed UI
- Current interaction techniques (GUI, TUI, manipulable UI)

The components within this framework can be summarized as in the following figure.



Hereafter several properties of actors and actuators were introduced. Interaction resources may have various properties like: shape, color, size, weight, width, height, material and texture. Surfaces may have the following properties: they may be solid, rigid, flexible or mobile. Depending on the data, the content may be public, private or semi-private.

The presentation closed with a list of requirements for the sensing & cognitive science communities (e.g., FGNet). These are:

- Detection of presence (arrival/departure) of instruments, surfaces and users
- Identification of instruments, surfaces, users, and their attributes and properties
- Identification of roles of instruments, surfaces, users
- 2D and 3D geometrical relationships between instruments, surfaces, users

The particular requirements are precision, stability, robustness to changes under natural conditions (light, heat, occlusions, etc.), recognition in real time (50ms human latency) and the ability to be reflexive (export performance at the software level).

4.3 Mohan Trivedi: “Distributed Video Arrays for tracking and activity analysis”

Prof. Mohan M. Trivedi presented his talk entitled "Distributed Video Arrays for tracking and activity analysis ". He is professor at the *University of California, San Diegeo, CVRR lab., Department of Electrical Engineering.*

Abstract

We are interested in developing intelligent environments which automatically capture and maintain an awareness of the events and activities taking place in these spaces. Such spaces

can be indoors, outdoors, or mobile. This is indeed a rather ambitious effort, especially when one considers the real-world challenges of providing real-time, reliable, and robust performance over the wide range of events and activities which can occur in these spaces. Novel multimodal sensory systems are required to realize useful intelligent spaces. Arrays of cameras and microphones distributed over the spatial (physically contiguous or otherwise) extent of these spaces will be at the front end of capturing the audio-visual signals associated with various static and dynamic features of the space and events. The intelligent environments will have to quickly transform the signal level abstraction into higher level semantic interpretation of the events and activities. We will present overview of our research directed towards the development of networks of video cameras, which support a wide range of tasks of intelligent environments. We will discuss real-time tracking of single or multiple people and on coordination of multiple cameras for capturing visual information on wide areas as well as selected areas for activity analysis, human body-movement tracking, and person recognition.

Presentation Outline:
The scope is the engineering of intelligent environments, which are decomposed into several subsystems. A short movie about the systems DIVA and MIA (Mobile Interactive Avatar), a campus tour guide, was presented.
Out of this movie the following questions were derived: What can not be done today?

- Cameras are only for viewing and recording
- Cameras have specific functions.
- Most cameras can not be used in an array.

Need for ubiquitous visions to capture and maintain awareness of dynamic events of variable spatio-temporal resolution and of multiple levels of abstraction.

Comments:

- Distributed sensor arrays.
- Face detection and localization.
- "Active" event based capture has lots of promise.
- "Attention" and "capture" have bi-directional relationships.
- Multiple disciplinary approach is essential.
- Relation requires development and evaluation in real-world situations.
- "Humans" may make reliable operation of a machine harder!

Intelligent Environments:
Environmental awareness can be separated into the so called static space awareness and the dynamic activity awareness. They can develop and maintain awareness of events. They can adapt both dynamic changes in their surroundings and they can interact in a natural, efficient and flexible manner. The results are tele-viewing, summarization and recall.

In this context the project DIVA and a few more were introduced in more detail. The systems can track and identify individuals on multiple levels. They can also do activity analysis on several type of interaction with active participants and present people in the room as well as with remote and future participants.

Multiple Abstractions and involved disciplines:
The following subsystems were involved in the presented project videos:

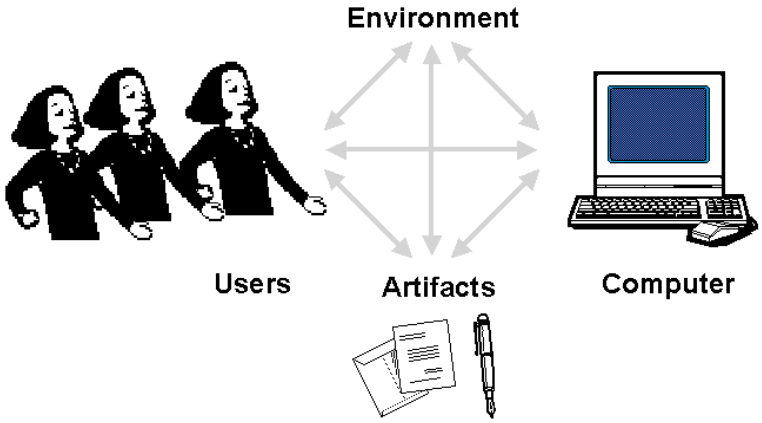
- Simultaneous 3D tracking of multiple blobs (project NOVA)
- Face recognition / orientation
- Capture of "interesting" things
- Activity summarization and recall
- Head tracking on PTZ (pan-tilt-zoom) camera
- Expression analysis and facial animation
- Making intelligent systems learn
- Video arrays for ubiquitous coverage (+thermal infrared)
- Body modeling and movement analysis system for posture and gait (10 segments in the MICASA project)
- Motion caption and evaluation of surveillance
- Televiewing and structural health monitoring
- Super bowl "crowd size estimator"
- Tracking, identification and activity analysis
- Distributed video networks and event based serving
- Context aware maps for situational awareness: "Human centered intelligent driving support system"

4.4 Michel Beadouin-Lafon: “Situated Interaction - creating interactive systems in context”

The third invited talk “Situated Interaction - creating interactive systems in context” was held by . Michel Beaudouin-Lafon from the *Université Paris-Sud.*

Abstract
Situated Interaction is an approach by which the design of an interactive system takes into account its context of use and takes advantage of the complementary aspects of humans and computers. This approach is meant to address 3 major challenges: the use of interactive systems by a wider audience for all aspects of everyday life and professional activities, the availability of a more and more diverse range of interactive devices, from cellphones and PDAs to immersive VR systems, and the ever increasing amounts of data that users need to cope with. In this talk I will introduce Situated Interaction and illustrate it through a number of example projects. I will then draw some conclusions about the use of recognition techniques for situated interfaces.

During the opening the meaning of the term interaction was briefly discussed. Interaction is a bi-directional way of communication between a user, which perceives, thinks and can act on the one hand and a computer, which stores, computes and has input- output functions on the other hand.
The new challenges that arose in the last years are based on the diversity of professional and non-professional users, different application areas such as professional, home use and entertainment, and interactive devices as for example PDA's, cellular phones or CAVES. Furthermore the scalability has an impact. This can be caused by the amount of people or users with different group sizes and the form of the data like documents or messages. Such a situated context is summarized in the following figure.



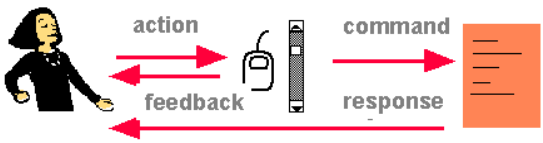
After this, the differences between the terms situations and context were discussed using theories of Suchman, Gibson and Mackay:

- During a situated action, which consists of perceiving, planning and acting in cognitive theories, plans can be revised during a particular situation.
- Ecological perception is active and extracts invariants, which affords an specific sort of information picking.
- In co-adaptive systems, users adapt to technologies as well as they adapt the technology be reinterpretation.

Context is the part of a situation that can be captured.
Then the diverse meanings between interaction and computation were shown. The fundamental assumption is that interaction is more powerful than an algorithm [Wegner]:

- Interactive system = open system
- Harness the power of the environment
- Rationalism vs. empiricism: models considered to be harmful

The challenges that arise from that are the design of interactive systems that work in unpredictable environments. The danger lies in the reduction of the environment, respectively the users to the obeyed algorithms and models.
The interaction paradigm were defined to be communication and instrumentalism. Communication can be: interaction (=exchanging messages), with a computer as partner, agents, avatars, natural language interfaces. Typically the communication requires a shared "code" on a cognitive level. Instrumentalism on the other hand can be interaction using a tool, can be with a computer as instrument, direct manipulation of interfaces. Instrumentalism requires a shared protocol on an action and perception level, and extends therewith human capabilities.
Instrumental interaction can be interpreted as mediation with objects of interest, which can be a reification process, where a command becomes an instrument, which is illustrated below:



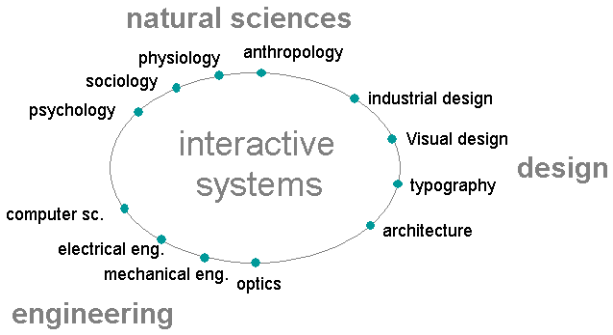
Recognition based interfaces are somewhere between communication and an instrument. They requires a model of possible input and have to manage ambiguity. Aspects of communication are recognition (= coding). The goal often is "natural" interaction and is not self-revealing. Aspects of instruments are capture devices and their feedback

(action/perception loop), which implies interaction. The question is: "What to recognize?" and "Why recognize?"

For "In Situ Computing" one has to understand the context of use by a study of the users, their artifacts and their environment. One has to further extract invariants by identifying the patterns of use. In addition to this, one has to take care of the design for minimal invariants: Power vs. simplicity as well as for the design for reinterpretation: Design principles.

All these theoretical aspects were recapitulated on several real world examples and related projects, which are Charade [Baudel], Caméléon [Mackay], A-Book[Mackay], CPN2000 [Beaudouin], InterLiving [Mackay et al.], VideoProbe - IST Interliving and MirrorSpace – IST Interliving.
The projects were introduced very briefly together with a special eye on the key design decisions, the results and their specific problems.

A comparison of the several introduced projects again demonstrates the multidisciplinary background of the approaches as depicted below:



After a short introduction of the Triangulation theory by Runkel and McGrath and related research strategies, the author showed a circle for the participatory design, which consists of observation, brainstorming, prototyping and evaluation. In this context a good trade-off between the power and simplicity of HCI has to be found.

The author concluded his presentation with considerations:

- To model or not to model (context vs. situation), what and how to model.
- Generality vs. specificness (power vs. simplicity).
- Situated interactions calls ad-hoc designs.
- Design methods and principles.
- Re-define the problem as well as the solution.

The key insight should be to fight the myth of perfection, which means, that systems can, and need not to recognize everything correctly. This leads to a design for incompleteness and ambiguity as well as a design for reinterpretation and unanticipated use.

4.5 Alan Johnston: “Dynamic Faces: Perception and Animation”

The last talk from Prof. Alan Johnston, *Department of Psychology* at the *University College London* was: “Dynamic Faces: Perception and Animation”.

Abstract
Most of the work on face perception has used static images of faces as stimuli but faces are dynamic channels of communication and we may recognise people by the way their faces move as well as their facial shape. In order to experiment with facial movement we need to be able to separate movement from shape. We achieved this by tracking the motion of points on the face and mapping that movement on to a standard 3D graphic head. Subjects are able to classify movements on the basis of identity and make categorical judgements such as the indicating the sex of the target performer at a level of performance greater than chance. We also investigated viewpoint dependence of facial movement. We found recognition of non-rigid, face-based motion to be more viewpoint invariant than rigid head movement, indicating object-based motion supports viewpoint generalisation whereas rigid head motion is encoded in a more view specific manner. I will also describe new work on facial animation that uses optic flow techniques to establish frame by frame correspondence rather than marker tracking and which delivers photorealistic performance driven animations.

First the several type of motion were introduced, which are: local motion, where local measurements are tied to locations. Then object motion, where motion descriptors are tied to objects. Third there can be object-based motion, where the change is tied to the parameters and constraints of an object's structural description. Fourth there are gestures, which are systematic dynamic patterns of object-based motion.

For moving faces this means, that facial movement is often gestural. Emotional expressions include dynamic change. Object-based motions involves configural change. Encoding of static faces may reflect these dynamic changes in faces.

One interesting arising question is. "Can we recognize people from the way their faces move?" Therefore photographic negatives were presented. This kind of manipulation does not change the position of features or the motion field.
For performance-driven facial animation was need the modelling of a face (creating a puppet). We need to track an actor to drive the model later (pulling the strings).
In a few interesting experiments it was demonstrated, that one is able to classify sequences on the basis of established categories such as gender. Especially the influence of the viewpoints was inspected at 0°, 15°, 30°, 45°, 60°, 75° and 90° by testing one unknown movement with one or more given motions. The recognition was almost worse by a major variation of the view angle.
Previous face modeling approaches often appeared to be synthetic. Complex polygonal 3D facial models or hand coded underlying muscle structures (3D mesh) were used. New approaches have a focus on being photo realistic. They use an image based representation. An automated generation from example video footage is used for this.
For the internal representation a simple form is used, which is a simple gray level intensity vector. With several of these vectors an avatar can be generated by a PCA (principle component analysis). One problem that arises is the blurriness of the so generated images.
Therefore an alternative, called the warped-based vectorization is introduced. Because the luminance is too blurred, the author considers each face in a sequence as a warp from the reference, which is called a flow field. Then these flow fields are vectorized. This warp based vectorisation solves the problem of blurriness, but it can't capture iconic changes. The combination of both approaches is the solution.

Several animated sequences with different avatars were demonstrated to impressively show the implemented techniques.

The presentation closed with a recapitulation of the major benefits of the presented approaches. These are photo realistic, automated generation of example video footage, no need for special equipment and low dimensional features. The presented technique can capture low dimensional content. It is real time implementable.

5 James Crowley: "Context Aware Observation of Human Activity"

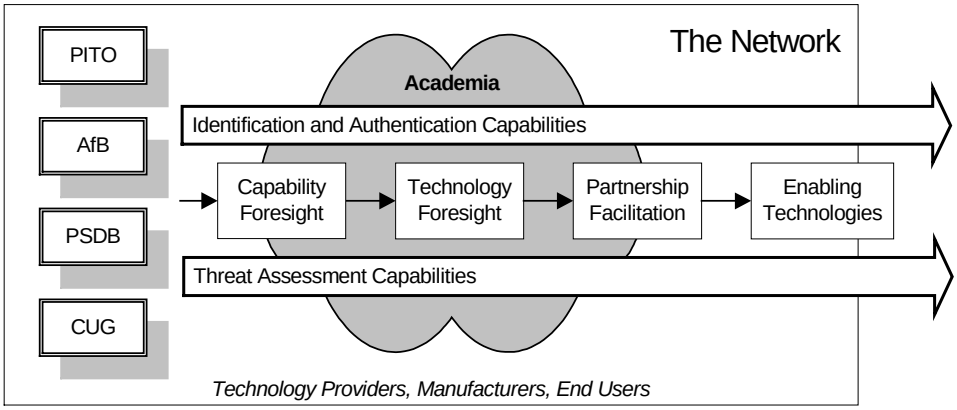
In addition to the invited speakers, the FG-net member Jim Crowley reported about recent related research activities.

Abstract
Human activity is extremely complex. Current technology allows us to handcraft real-time perception systems for a specific perceptual task. However, such an approach is inadequate for building systems that accommodate the variety that is typical of human environments. In this paper we define a framework for context aware observation of human activity. A context in this framework is defined as a network of situations. A situation network is interpreted as a specification for a federation of processes to observe humans and their actions.
We present a process-based software architecture for building systems for observing activity. We discuss methods for building systems using this framework. The framework and methods are illustrated with examples from observation of human activity in an "Augmented Meeting Environment".

6 James Ferryman: " Video-based Threat Assessment: ViTAB Network and related EU Projects "

Within this agenda item, the presenter James Ferryman from Reading University gave an overview over the objectives of the project ViTAB. ViTAB is an abbreviation for Video Based Threat Assessment. A short overview over other related EC funded projects followed:

- Overview of the ViTAB Network
  - To encourage maturation of image- and video-interpretation technologies most likely to reduce level of crime
  - To influence research activity within image processing, computer vision, machine learning and signal processing communities
  - The emphasis is on technologies that
    - improve effectiveness of CCTV
    - enable identification and authentication capability
  - The network combines academics, technology providers and police-service end-users
- Impact of Network
  - Industry and Users



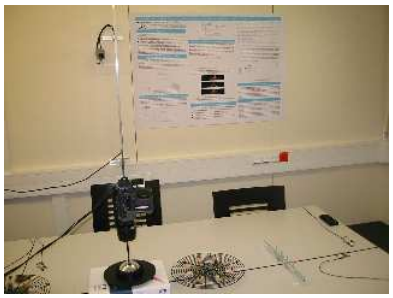
- Academia

The requirement to develop key policing and crime reduction capabilities in the areas of threat assessment, identification and authentication are the two primary drivers of the structuring activities within the ViTAB network, which is illustrated in the figure above. It further shows the roadmap for the technology partner, manufacturers and the end users.
Related to this network are the two following EC projects within the 6th framework programme:

- SAFEE: The vision: The construction of an advanced aircraft security system designed to operate during on-board terrorist threat scenarios.AVITRACK: "Aircrafts, Vehicles & Individuals Categorisation & Tracking for apRon's Activity Model Interpretation & ChecK"

7 James Ferryman: "PETS workshops"

In his second talk James Ferryman from the University of Reading presented an overview over the last performance evaluation workshop PETS-ICVS 2003 in Graz.
In this context the data, which was planned and prepared by the FGnet community were reviewed. The data is now available on a set of 2 DVDs and is also online available. The volumes contain data to test "Observing People Interacting in Meetings". Examples of pictures of the gathered material are given below.



Furthermore the next workshop called VS-PETS, which will be held in October 2003 in Nice in conjunction with the ICCV, together with its aims was introduced. Hereafter several possibilities for a PETS workshop 2004 were discussed, see next section.

8 Data Acquisition for Benchmarking

One of the objectives within the FGNet is the generation of resources and to encourage other researchers to test their algorithms under well defined constraints. Therefore, in conjunction with this workshop a track again was dedicated to define some training and test sets to benchmark recent recognition systems for sub-tasks like face localization, recognition of facial expressions, recognition of face/hand gestures, estimation of face/head directions and recognition of actions.
It was decided to acquire benchmarking datasets, which are related to the topic "Human Machine Interaction". In contrast to last years benchmarking strategies, the next data set will most likely not be assembled for the next PETS workshop. The reason for this is the fact, that data for human machine interaction would not satisfy the needs for the initial topic of the PETS workshop, which stands for Performance Evaluation on Tracking and Surveillance.
It was therefore decided to set up a separate workshop in conjunction with on of the major conferences for the vision community. This could be the CVPR'04 in Washington, the ECCV'04 in Prague, or most likely the ICPR 2004 in Cambridge.

The latest conclusion about the content of possible datasets was to have one part to benchmark systems that can solve a pointing task. One or two cameras behind a projection plane record several persons, pointing to defined spots within the plane. The second part is considered to evaluate systems that recognize the gaze of persons. The data collection mechanism could be similar to the first one. Furthermore a third dataset, namely one containing pointing gestures recorded via a head mounted camera were discussed.

9 Foresight Visions

After all workshop contributions have been presented, the participants were divided into two groups, A and B. The goal was to find possible scenarios which intersect with the topic “Human Machine Interaction”. The results of the brainstorming process is integrated in a second step.

9.1 Group A

The members of group A were: Ipke Wachsmuth, Michelle Beadon-Lafon, Gerhard Rigoll, James Crowley, Agnes Just and James Ferryman. The following list is the unfiltered output of the brainstorming process.

How will HCI-based FGnet methods (Perception for Interaction) develop over the next few years?

Definition:

- Employing face+gesture methods for interaction with machines
- Visual observation of humans
- Multimodal input and output

Application domains

- Computer games
- Disabled – sign language
- Human-robot communication
- Communication in mobile environments
- Smart rooms
- Information kiosks

Decision support systems via CSCW – CMC (computer mediated communication)

- Command and control
- Video communications – videoconferencing, video telephone – personalised services
- Training, tutoring
- Television – media-metrics; emotion-aligned content; access control
    - Improved VCR interface

What techniques will be required?

- Recognition of emotion
- Detect and track hands and fingers and grasped objects
- Detect and track faces and components
- Recognition of pointing gestures
- Estimate face orientation
- Estimate gaze orientation
- Detecting speech acts
- Fusion – integration (at multiple scales, levels of abstraction, temporally …)
- Non-keyboard text input
- Touch-sensitive surfaces (integrated)
- …

What datasets are to be collected over the next 6-12 months?

- Face orientation and gaze direction (measuring)
- Skill task requiring hand/object manipulation (for training/tutoring)
- Virtual keyboard
- Tabletless tablet – capturing pen input (2 versions – paper & whiteboard)
- Playing a musical instrument
- Pointing gestures
- Online – positioning icons by finger tracking

Topic of final dissemination meeting?

- Possible performance evaluation in one or two of the following fields:
    - Face orientation and gaze direction
    - Virtual keyboard
    - Positioning icons by finger tracking
    - Playing a music instrument
    - Capturing pen input on whiteboard
    - Pointing gestures

9.2 Group B

The members of the second group were: Alan Johnston, Andreas Lanitis, Tim Cootes, Joelle Coutaz, Thomas Moeslund and Frank Wallhoff. The following list is the unfiltered output of the brainstorming process.

Roadmaps/Scenarios:

- Study Scenarios
- Mobile Users
  - Telephones
  - Portable Computers /PDAs
- Privacy Aspects
- User Demands
- Active Walls
- Types of user interfaces depending on the location
- Avatars and artificial "Partners"
- Intuitiveness
- Involved Instruments
- Involved Technologies
- Handicaped/Disabled Users
- Safety
- Hands free computing
- Criminal Behaviour / Safety
- Computer mediated Communication
- Situated Information Portals
- Identification of user
- Accessabilty of resources
- Multi-Level Integration
- Adapt system to users needs NOT vice versa
- Training Robots, machines in general
- Controlling Machines
- Augmented/Mixed Reality
- Automating Interaction
- Head Mounted Displays
- Access Services by gestures
- Entertainment / Controlling Games
- Educational Systems
- Implicit Interaction
- Invariance of cultural aspects (language...)

Techniques:

- Improving sensors (resolution, compression, speed...)
- Facial Expression Recognition
- Face Identification
- Model Human behaviour (pieces / global)
- General Body part detection
- Action Segmentation/Recognition
- Speech Recognition/Speaker Identification
- Autocalibration

- Ageing Evolution
- wearability
- low power consumption
- low level Segmentation
- (hand-) gesture recognition
- find vocabulary/"information atoms"
- Role and situation Recognition
- Scalability
- Phsychological Aspects
- Broaden up scientific spectrum
- Infrared Light
- Intelligent Cameras
- Reduce computational power
- 3D Scanners
- Distributed systems

Databases that could be relevant to test the techniques above:

- In-Car Devices (to test multimodal systems)
- Sequences of interacting people (to learn about the (natural) way they-also non verbal-communicate
- Gestures for a wearable computer for robustness against varying environmental constraints (light (direction, intensity, temperature, background, ...)
- Aging (large scale / short term)

9.3 Final Roadmap after Integration and Filtering

In this final phase the output of both groups was integrated by taking the list of the first group, and sorting in the entries of the second group. Double, and unprecise entries were discarded. The following definitions are used: (S)=Short term (< 5 years), (M)=Medium-term (5-7 years), (L)=Long-term (10-20 years)

How will HCI-based FGnet methods (Perception for Interaction) develop over the next few years?

Definition:

- Employing face+gesture methods for interaction with machines
- Visual observation of humans
- Multimodal input and output

Application domains:

- Entertainment and Computer games (S)
- Disabled, impaired (S)
- Sign language (S)
- Human-robot communication and control and training (M)
- Avatars and Personal Assistant (L)
- Interaction (and communication) in mobile environments (S)

- Hands free interaction (M)
- Augmented spaces (active walls, smart rooms) (L)
- Augmented objects (L)
- Information kiosks (M)
- Decision support systems via CSCW – CMC (computer mediated communication) (L)
- Command and control, control of heavy equipment (S)
- Video communications – videoconferencing, video telephone – personalized services (S)
- Education, Training and tutoring of people (M)
- Television – media-metrics; emotion-aligned content; access control (M)
- Improved VTR interface (M)

Virtual and mixed reality: determining new forms of interaction through observation of object manipulation

Ethical concerns, including privacy, trust, possession

What techniques will be required?

- Recognition of emotion
- Facial Expression recognition
- Face Identification
- Detect and track hands, fingers and grasped objects
- Detect and track body parts
- Detect and track faces and components
- Recognition of pointing gestures
- Estimation of face orientation
- Estimation of gaze orientation
- Detecting, parsing and recognizing actions
- Detecting speech events
- Fusion–multimodal integration (at multiple scales, levels of abstraction, temporally)
- Non-keyboard text input
- Touch-sensitive surfaces (integrated)
- Modeling human gestural behaviour
- Synthesis of human gestural behaviour
- Role and situation recognition
- 3D Scanning

Criteria:

- Auto-calibration
- Robust to variations in lighting, view angle, in scale and time
- Real time
- Scalability in terms of number of entities and sensors
- Acceptability
- Graceful degradation
- Computational Complexity
- Precision
- Stability

What datasets are to be collected over the next 6-12 months?

- Face orientation and gaze direction (measuring)
- Skill task requiring hand/object manipulation (for training/tutoring)
- Virtual keyboard
- Tabletless tablet – capturing pen input (2 versions – paper & whiteboard)
- Playing a musical instrument
- Pointing gestures
- Online – positioning icons by finger tracking
- In-Car Devices (to test multimodal systems)
- Sequences of interacting people (to learn about the (natural) way they-also non verbal-communicate
- Gestures for a wearable computer for robustness against varying environmental constraints (light (direction, intensity, temperature, background, ...)
- Aging (large scale / short term)

Topic of final dissemination meeting?

- Possible performance evaluation in one or two of the following fields:
    - Face orientation and gaze direction
    - Virtual keyboard
    - Positioning icons by finger tracking
    - Playing a music instrument
    - Capturing pen input on whiteboard
    - Pointing gestures

10 Summary and Conclusions

During the third 2-day lasting workshop, the state-of-the-art in the topic "Human Machine Interaction" and the non-technical issues resulting from the deployment of this technology have been demonstrated by the invited speakers and discussed by the workshop participants.

Finally, a large variety of foresight visions have been defined and elaborated by the two groups formed at the final afternoon of the workshop. These should be helpful in order to indicate development roadmaps and opportunities for the technology in the short (<5 years), the medium (5-7 years) and long term (>10 years), respectively to estimate the difficulty, economic or social payoff and a possible market size.

11 Acknowledgment

This workshop was also sponsored and announced by the IEE and IEEE-Cyprus section.

12 References

Sowa, T., & Wachsmuth, I. **"Interpretation of Shape-Related Iconic Gestures in Virtual Environments."** In I. Wachsmuth & T. Sowa (Eds.), Gesture and sign language in human-computer interaction. Berlin: Springer, 2002.

Coutaz, J. and Lachenal, C. and Dupuy-Chessa, S. **"Ontology for Multi-surface Interaction",** To appear at Interact 2003, Zurich, IOS Press, 2-5 September 2003.

Trivedi, M. M. and Huang, K. and Mikic, I. **"Dynamic Context Capture using Distributed Video Arrays for Intelligent Environments",** IEEE Transactions on Systems, Man and Cybernetics, Special Issue on Ambient Intelligence, Submitted October 2003.

M. Beaudouin-Lafon and W. Mackay. **"Prototyping Development and Tools"**. Human Computer Interaction Handbook, J.A. Jacko and A. Sears (eds). Lawrence Erlbaum Associates, 2002

Hill, H. and Johnston, A. **"Categorizing sex and identity from the biological motion of faces. Current Biology"**, 11, 880-885, 2001

Crowley, J. L. and Reignier, P. **"An Architecture for Context Aware Observation of Human Activity"**, Workshop on Computer Vision System Control Architectures (VSCA), March 2003.

Appendix I -Final Programme

Thursday, 28. August 2003:

9:00	Arrival - Welcome Discussion
9:30	Frank Wallhoff and Andreas Lanitis: Opening and Welcome from the workshop organizers and the local host
9:45	Tim Cootes: Message from the project-coordinator, Overview over FGNet Progresses and Status Report
10:00	Gerhard Rigoll: Introduction to workshop topic „Human Machine Interaction“ Schedule-Background-Goals
10:15	Coffee break
10:45	Ipke Wachsmuth: „Embodied Communication“
11:45	James Ferryman: „Video-based Threat Assessment: ViTAB Network and EU projects“
12:30	Lunch at Hotel-Restaurant
14:00	Joëlle Coutaz: „Distributed User Interfaces and Multi-surface Interaction“
15:00	Mohan Trivedi: „Distributed Video Arrays for tracking and activity analysis“
16:00	Coffee break
16:15	James Ferryman: Feedback on PETS-ICVS, plannings for VS-PETS at ICCV (Nice,October 11-12), and future plans for PETS 2004 (finalising and collecting datasets)
16:45	Michel Beaudouin-Lafon: „Situated Interaction - Creating Interactive Systems in Context“
18:00	Break
18:15	FGnet Management Meeting
20:00	Dinner at the St. Raphael's Restaurant

Friday, 29. August 2003:

9:00	Arrival Discussion
9:15	Alan Johnston: „Dynamic Faces: Perception and Animation“
10:30	Coffee break
10:45	James Crowley: „Context Aware Observation of Human Activity“
12:00	Foresight-Visions: First discussions, Splitting into 2-3 groups
12:30	Lunch at Hotel-Restaurant
14:00	Defining Road-Maps (groupwise brainstorming)
16:00	Coffee break
16:15	Integration and Filtering of individual group results
18:45	Workshop summary & wrap-up (G. Rigoll)
18:00	Break
18:15	FGnet Management Meeting
20:00	Dinner

Appendix II -List of Participants

NAME	INSTITUTION	COUNTRY	ROLE
James Ferryman	University of Reading	UK	FGNet Partner
Tim Cootes	University of Manchester	UK	FGNet Partner
James L. Crowley	INRIA Rhône Alpes	F	FGNet Partner
Joelle Coutaz	INRIA Rhône Alpes	F	Invited Speaker
Ipke Wachsmuth	University of Bielefled	D	Invited Speaker
Mohan Trivedi	University of California, CVRR	US	Invited Speaker
Michael Beadon-Lafon	University of Paris South	F	Invited Speaker
Andreas Lanitis	Cyprus College	CY	FGNet Partner
Agnes Just	IDIAP	CH	FGNet Partner
Thomas Moeslund	University of Aalborg	DK	FGNet Partner
Alan Johnston	University College London	UK	Invited Speaker
Gerhard Rigoll	Technische Universität München	D	FGNet Partner
Frank Wallhoff	Technische Universität München	D	FGNet Partner
Chris Christodoluou	Birkbeck, University of London	UK	Guest
Chris Constantinides	PHD-Student	CY	Guest

Participants of the third FGNet workshop

Group photo of the participants from left to right: James Ferryman, Agnes Just, Thomas Moeslund, Alan Johnston, Tim Cootes, Andreas Lanitis, Gerhard Rigoll, Frank Wallhoff, Ipke Wachsmuth, Michel Beadon-Lafon, Joelle Coutaz, James Crowley.