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	*15-1-2003*
Actual date of delivery	*27-3- 2003*
Deliverable number	*D2.2*
Deliverable name	Workshop 2 Report
Type	*Report*
Status & version	*Version 1.0*
Number of pages	*14*
WP contributing to the deliverable	*WP2*
WP / Task responsible	*WP2 Foresight Workshop*
Other contributors	-
Author(s)	Frank Wallhoff
EC Project Officer	*Phillipe Gelin*
Keywords	*Face Recognition, Gesture Recognition, Foresight Report*
Abstract (for dissemination)	The second FGNet foresight workshop was held at IDIAP in Martigny from 12 - 13 September 2002, with the subject of “Interacting People”. This document describes the workshop and the a series of scenarios for suggesting computer vision may help people interact in the future.

FGNet - 2nd Foresight Report

Date of preparation: 27 Mar 2003

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 second FGNet foresight workshop was hosted by Sebastien Marcel at IDIAP in Martigny from 12. - 13. September 2002. This document reports about the content and outcome of this second workshop. The outline of this report is as follows: First, a brief introduction into the workshop topic “Interacting People” is given. The subsequent section contains summaries of the talks presented by the invited speakers and related projects of the network members. The next section describes the foresight visions that were defined by two different working groups at 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 Introduction into the workshop topic “Interacting People”

It was decided by the workshop organizers that each foresight workshop should be associated with a specific topic from the field of face & gesture recognition. Talks and discussions at the workshop should be mainly centered around, but not strictly limited to this workshop topic.

This year’s workshop topic was selected to be “Interacting People”. The topic is narrowly restricted to people that interact with each other. This topic covers scenarios such as:

- Face to face meetings
- Video conferences
- Seminars and lectures
- Negotiations
- Sign language conversations

One of the reasons, why this workshop topic has been chosen, is the fact that “Interacting People” is a new and actual topic that is currently addressed in several recently started EC projects, some of them are mentioned in more detail in this report later in section 4 and 5. Another reason for this topic is also the fact that it involves many interesting FG-disciplines, such as e.g.:

- Surveillance
- Tracking
- Gestures
- Action recognition
- Facial expressions
- Emotion recognition
- Person identification
- Audio-visual speech recognition

Additionally, this scenario is that much complex that the recognition results are in most cases not of “elementary nature”, such as e.g. recognition of single gestures or simple visual patterns, but often are of “higher semantic nature”, involving the identification of higher level events, such as e.g.:

- Decisions in meetings
- Agreement or disagreement in negotiations
- Identification of presentations in conferences

Thus, this topic is highly demanding and requires sophisticated algorithms involving low level signal and image processing methods as well as high level interpretation of the recognition results.

3 Summary of the invited talks

Three invited talks were presented at the workshop, which are briefly summarized separately in the following paragraphs. They are all closely related to the theme of the workshop.

3.1 Simon Rowe: “Video Conferencing, Face Modeling, Gaze Tracking”

The first invited talk was presented by Dr. Simon Rowe, group leader of the *Visual Communication Technology Group* at *Canon Research Center Europe* (CRCE), entitled “Video Conferencing, Face Modeling, Gaze Tracking”.

After a short introduction of the Canon research group, the motivation for improvements of video conferencing was presented. Video conferences can save time and traveling costs. They furthermore can bring together people, which would not be able to meet in the real world. The major problem is that traditional video conferencing tools have restrictions, so that conference attendees may be confused by the unnatural conference situation caused by the different locations of the camera and the displays. So questions arise like “Where to speak to?” or “Who speaks to whom?”. These unsatisfactory facts make it desirable to derive avatars from the conferees. The idea is to model synthetic, realistic and natural avatars, so that people which look at each other on the corresponding conference image, see that they look at each other, not having to look straight forward into the camera. The other participants can then also see that these people look at each other and not at them. Goal is to have several asymmetric animations derived from the same asymmetric arrangements but share the same semantics.

The above mentioned problem contains several sub-problems like visual perception, static photographs, 3D photos and 3D head models. However, the plan to derive a virtual meeting environment brings up two interesting major topics, which are face modeling and gaze tracking.

Today’s 2D/3D face models can have resolutions from 120x120 up to 4000x3000 pixels, depending on the application. For this reason the work of Tim Cootes, which synthesizes face images from given 2D images has been introduced. An alternative approach of Vetter and Blanz using 3D models has also been mentioned. However, the used models have to cope with all categories of faces such as gender, ages and races. They furthermore have to produce realistic views.

The second aspect was the tracking of the gaze, which aims to find out who is looking at what or whom. Here two different approaches have been mentioned which are active and passive trackers in general. The active one use the location of the eyes and the position of the pupils, which have been recorded with infrared cameras which is very expensive, but stable. The passive approaches find the center of the eyes and the view direction of the pupil and derive the gaze information by these correlations, which is computationally expensive on the other hand.

In summary, this talk impressively shows that recent systems for “interacting people” have to adapt to the customers needs, which means that man machine interaction has to be natural.

3.2 Christer Fernström: “Perception of Human Activity in Knowledge Management”

Christer Fernström, Area Manager in the *Coordination Technologies Group* at the *Xerox Research Center Euro* (XRCE) located in Grenoble, presented the second invited talk which is about “Perceptions of Human Activity in Knowledge Management”.

The introduction opens with an overview over the Xerox company. Then the company’s mission was pointed out:

- Bridging the gap between paper documents and electronic documents,
- “Green Button” for zero effort (everything works automatically), and
- Generation of multifunction and intelligent copiers
- Intelligent knowledge sharing

To solve the problems stated above it is necessary to detect the user activities sorted by users and service technicians. It is further important to retrieve or search for information no matter in what document type they are stored. Xerox has invented several technologies for this task. One of them is the so called “askOnce” meta-search engine to support knowledge management, with a particular focus on the support of efficient diffusion and sharing of information in large organizations among various people.

An extension to this approach is the so called “Knowledge Pump” recommender engine, which works on user profiles and therefore also on a semantic layer. The idea is that one person stores information with context, so that a second person with identical interests has also access to this preprocessed data from the first person.

After an introduction of this complex information sharing and retrieval system, a few applications together with real world scenarios were introduced. One idea is the automated storage and archiving of information into a common database, during just copying documents on a smart copier. Another scenario is the so called “community wall”, which is a special form of a huge interactive touch screen, also called ambient displays. This device represents a smart user interface, that is directly picking up the idea of observing interacting people. The community wall is already actively used at Xerox and deploys various FG-methods, such as face detection, gaze detection and action recognition in order to identify relevant group activities in rest areas and discussion corners at Xerox facilities. It is therefore an impressive example in order to apply advanced techniques for analyzing people interactions to improve the efficiency of the personals cooperation in large companies.

The presentation closes with an overview of outside the office applications, which are touring, shopping, and information exchange.

3.3 Volker Krüger: "Human Identification at a Distance – Human ID"

The last talk from Volker Krueger, *Center for Automation Research* (CFAR) at the *Institute for Advanced Computer Studies* at the *University of Maryland* was: "Human Identification at a Distance – Human ID".

The focus of many researchers lies in the development of robust methods for person identification from video based on gait. In this context a hidden Markov model (HMM)-based approach to recognize humans using gait was proposed. The suitability of HMM for the gait problem stems from its ability to effectively encode structural as well as transitional features that constitute the gait of an individual. Gait is a spatial-temporal phenomenon and the HMM is well suited to capture the inter-dependencies that exist across frames during a walk cycle. Moreover, the statistical nature of the HMM improves overall robustness of gait representation and recognition. The observation sequence for the HMM is derived by extracting the outer contour of the body during a walk cycle. A novel methodology is here used for data dimensionality reduction and a HMM is trained to learn the gait of each person in the database. Recognition results are presented. The performance of the proposed method when tested on real video sequences is found to be quite good compared to alternative approaches.

However, an extension to this first approach, the 3D gait based Human ID project was started. The aim of the new project is to develop a 3D approach, which recovers the 3D characterics of both body shape and body motion using 3D-models. An approach to track human activities in color video was also presented. The human body, decomposed into truncated cones and ellipsoids is used. The body parts are organized as a tree with an ordered chain structure to provide the kinematic model of the limbs. The motions of the limbs are the rotations at the joints, and are represented using the relative rotation between local coordinate systems angles. The problems of motion tracking and estimation are posed as a nonlinear state estimation problem. The measurements are computed using the outputs of 3D shape-encoded filters which extract the boundary gradient information of the body image. The nonlinear state estimation is performed by solving the Zakai equation using branching and the particle propagation method.

The presentation closed with an ice skating sequence, where the body of the skating person was tracked very precisely using the 3D model.

4 Introduction of the IST FAME project

The speaker James Crowley informed the workshop participants about the major ideas in the running IST FAME project. FAME is an acronym for "Facility Agent for Multi Cultural Exchange". The project consortium consists of members from industry, research and education.

The major goals of this project are to facilitate human-human communication through multimodal interaction including vision, speech and object manipulation, provide appropriate information relevant to the context, and enable production/manipulation of data-blending by electronic and physical representation. Another goal is to construct an intelligent agent to facilitate communication among people at separate locations from different cultures who collaborate on a common problem.

The major challenges within this project will be automatic perception of human action and understanding of human free dialog between people from different cultures. The consortium will construct an information butler, which demonstrates context awareness in a problem-solving scenario using computer vision, speech and dialog modeling. The result will be a class of integrated tools for computer enhanced human-human communication. A public demonstration is planned for the Barcelona Cultural Fair 2004.

In addition to the project objectives a short overview of possible ontologies in the context of the observation of human activities was presented. This can be property, entity, relation, role, situation and context.

5 IST project M4 and IDIAP's Smart Meeting Room

Within this agenda item, the presenters Daniel Gatica-Perez and Darren Moore from IDIAP first gave an overview over the objectives of the so called IST project M4, which stands for MultiModal Meeting Manager. A guided tour through the IDIAP smart meeting room followed.

The M4 project is about processing meetings held in a room equipped with multimodal sensors. The overall objective is the construction of a demonstration system to enable the offline structuring, browsing and querying of an archive of meetings. The project will include the design, collection and annotation of a multimodal meeting database, the processing of audio/video streams and the integration and structuring of these streams using the outputs of various recognizers and analyzers. The assumption of the availability of textual side information (e.g., an agenda) is made, which enables the application of some useful constraints. The expected results of the project include a demonstrator system, and advances in models and algorithms for multimodal recognition, integration and information access.

At the end of the introduction of the project, a short walk through the meeting room was arranged where the participants were split into two groups. The room consists of a well defined geometry. In more detail this means there is an equally lighted row of tables with equally distanced chairs. In one back of the room the recording technique is located, while in the other end of the room a white board is placed. On both of the transversal walls a camera is located to record the meeting attendees. A third camera records the action at the whiteboard.

In addition to this video equipment the room has several setups of microphone arrays. Meetings can be recorded directly in high quality with up to 3 color video and 16 audio channels.

6 Data Acquisition for PETS at ICVS

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 the workshop a track 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. These data will be disseminated for the IEEE PETS workshop in conjunction with the IVCS 2003. PETS stands for Performance Evaluation of Tracking and Surveillance systems. The theme of the workshop is "observing people interacting in meetings". For this purpose several scenarios with the above mentioned content are recorded in a typical meeting room atmosphere with 6 interacting people.

Therefore a set of actions and gestures was defined here, i.e. sit down, get up, writing, going to the board, talking, raising hand, nodding, shaking head, yawning, and laughing. In addition to this a list of facial expressions was considered like smile, laugh, angry, and neutral. The detailed outcome of the discussions together with the material itself are available online at http://www.cvg.cs.rdg.ac.uk/PETS-ICVS/pets-icvs-db-spec.html. It was agreed to make the recordings at the IDIAP meeting room introduced above.

7 Foresight Visions

After all workshop contributions have been presented, the participants were divided into two groups. The aim was to find possible scenarios which intersect with the topic “Interacting People”.

7.1 Group I

Definitions: (S)=Short term (< 5 years), (M)=Medium-term (5-7 years), (L)=Long-term (10-20 years)

- Observing crowds
    - hooligan detection (L)
    - leader detection (L)
    - crush area prediction (M)
- Face-to-face meetings
    - agreement/disagreement detection (L)
    - generation of minutes (M)
    - presentation detection (S)
    - annotation of entire section (L)
    - task assignments (L)
    - votings (S)
    - decisions (S)

    - virtual chairman (L)
    - realtime meeting content monitoring (e.g. for simultaneous multiple meeting attendance) (L)
- Video conferences
    - Physical agents (L)
    - Virtual agents (M)
    - Automatic cameraman (S)
- Seminars & interviews
    - Attention level detection (M)
    - Participation detection (S)
    - Conflict detection (L)
    - Cheating detection (L)
- Community wall, discussion corners
    - Detection of discussions (M)
    - Attention focus (S)
- Operating theaters (errors in surgery)
    - Online error detection & prevention (L)
    - Concentration detection (M)
- Games with interacting people
    - Automatic referee (L)
    - Game control (M)
    -

Possible Scenarios for data acquisition for benchmarking

- Face-to-face discussion
- Face orientation database
- Group of 2 or 3 people planning a route by looking at a map
- Meeting/Minutes recording
- Interaction recording in commercial environments (salesman/customer)
- Behavior analysis in public areas
- Intelligent camera man (record person speaking & person spoken to)
- Sports (fault detection)

7.2 Group II

Labels: (difficulty, economic or social payoff and market size, relevance to FG net) (1 to N where N is highest) [The labeling of the topics was not finished in time]

- Mobile Phone communication
    - Gesture understanding for communicating with the deaf (4, 3, 1)
    - Context aware telephones (2, 2, 1)
- Sports
    - Online statistics of game activity (10, 15, 4)
    - Game play archiving (10, 5, 2)
    - Penalties and referee abuses. (15, 2, 1)
    - Referee support (15, 2, 1)
- Entertainment
    - Games (1-100, 1-100, ??)

    - Television monitoring: Mediametry (2, 4, 0)
    - Interactive TV
    - Data base retrieval for movies and television productions.
- Product evaluation. (packaging, instruction)
- Education
    - Remote learning
    - Feedback to speaker of Audience Attention
- Security
    - Airport Security
    - Metro station: Crime detection,
    - parking lots: Crime detection
    - Shopping malls: Theft detection, suspicious behaviors
- Commercial Environment.
    - Profiling by observing attention
        - Keeping track of customer interests
        - (including across multiple stores)
    - Sales clerk and customer interaction.
        - Improve customer satisfaction
        - Are customers leaving satisfied with service
        - improve productivity of the store
    - Best practice learning
- Office Work Environment.
    - Enhance productivity of individuals
    - enhance productivity of groups
    - Providing information that is relevant to a meeting
    - remote working.
    - Call centers
    - Collaborative Work Tools (mediaspace, etc)
        - Privacy protection
        - providing appropriate information
        - recording minutes
- Meeting Management
    - Recording minutes (5, 5, 10)
    - presenting appropriate information
- Teleconferencing
    - Subproblems : Intelligent Camera man for camera direction, camera selection, view selection
    - (context aware video conferencing).
    - Presenting who is looking at what.
- Workpractice Analysis (10, 5, 10)
- Medicine
    - Surgeon and assistant
    - Remote surgery
    - Diagonsis of childhood behavioural problems

8 Planning of the 3rd Foresight Workshop

After a short discussion, the participants voted for Cyprus as the location for the third workshop. Therefore the next host will be the Cyprus College represented by Andreas Lanitis. The 28. and 29. August were chosen for the meeting. The topic will be “Human Machine Interaction”.

9 Summary and Conclusions

During this second 2-day lasting workshop, the state-of-the-art in the topic “Meeting People” 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 different 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 medium (5-7 years) and long term (>10 years), respectively to estimate the difficulty, economic or social payoff and a possible market size.

Appendix I Final Programme

Thursday, 12. September 2002:

9:00-9:30	Arrival - Discussion
9:30-9:45	Welcome from the local host (S. Marcel) General Information on workshop venue and local environment
9:45-10:00	Message from the coordinators (T. Cootes/C. Taylor)
10:00-10:15	Message from the workshop organizer (G. Rigoll) Introduction to major workshop topic "FG for Interacting People" Schedule - Background - Goals
10:15-10:45	Coffee Break
10:45-11:30	Invited Speaker: Simon Rowe, Canon Research Europe "Video Conferencing, Face modeling and Gaze tracking"
11:30-12:00	Discussion
12:00-14:00	Lunch at Restaurant
14:00-14:30	"PETS workshop at ICVS" (J. Ferryman/J. Crowley)
14:30-15:15	Daniel Gatica Perez and Darren Moore IDIAP's Smart Meeting Room and IST project (M4)
15:15-15:45	Coffee Break
15:45-16:15	Guided Tour through the room with recording example
16:15-16:30	Discussion
16:30-17:30	Conclusions for workshop day I
17:30	Return to Hotel
18:00	Dinner at Restaurant

Friday, 13. September 2002:

9:30-10:00	Arrival
10:00-10:45	Invited Speaker: Christer Fernström (Xerox): "Perception of Human Activities in Knowledge Management"
10:45-11:15	Discussion
11:15-11:45	Coffee Break
11:45-12:15	James L. Crowley, INRIA: "Introduction of the IST FAME project"
12:15-13:00	Lunch at IDIAP (buffet of various sandwiches)
13:00-13:45	FGNet: data acquisition for PETS
13:45-14:15	Invited Speaker: Volker Krüger, University of Maryland
14:15-14:30	Discussion
14:00-15:00	Foresight Visions (2 groups)
15:00-16:15	Discussions on Foresight visions, PETS data, and date and location for 3rd meeting
16:15-16:30	Workshop summary & wrap-up (G. Rigoll)
16:30	Return to Hotel
18:00	Dinner at the hotel

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
Christa Fernström	Xerox Research Center Europe	SW	Invited Speaker
Sadaoki Furui	Tokyo Institute of Technology	J	Invited Participant
Daniel Gatica-Perez	IDIAP	CH	Speaker
Volker Krüger	University of Maryland	USA	Invited Speaker
Andreas Lanitis	Cyprus College	GR	FGNet Partner
Sebastien Marcel	IDIAP	CH	FGNet Partner
Thomas Moeslund	University of Aalborg	DK	FGNet Partner
Darren Moore	IDIAP	CH	Speaker
Gerhard Rigoll	Technische Universität München	D	FGNet Partner
Simon Rowe	Canon Research	UK	Invited Speaker
Frank Wallhoff	Technische Universität München	D	FGNet Partner

Participants of the second FGNet workshop